Vol. 63, No. 4 July, 1956

Psychological Review

THEODORE M. NEWCOMB, Editor
UNIVERSITY OF MICHIGAN
Lorraine Bouthilet, Managing Editor

CONTENTS

Walter Samuel Hunter: 1889-1954 ..... J. Mc. V. HUNT 213

On Thought: The Extrinsic Theory ..... EUGENE GALANTER AND MURRAY GERSTENHABER 218

A New Theory of Color Vision ..... WILLIAM A. SHAW 228

A Field Theory of Extinction and Spontaneous Recovery ..... KOH JOHDAI 243

Mediation Variables in Scientific Theory ..... WILLIAM W. ROZEBOOM 249

The Contribution of John B. Watson ..... GUSTAV BERGMANN 265

PUBLISHED BIMONTHLY BY THE
AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

CONSULTING EDITORS

Solomon E. Asch	Robert B. MacLeod
Robert R. Blake	David C. McClelland
Stuart W. Cook	George A. Miller
Clyde H. Coombs	Gardner Murphy
Leon Festinger	Oscar Oeser
W. R. Garner	Carl Pfaffmann
James J. Gibson	Carroll C. Pratt
D. O. Hebb	John P. Seward
Harry Helson	David Shakow
E. R. Hilgard	Richard L. Solomon
Carl I. Hovland	Eliot Stellar
E. Lowell Kelly	S. S. Stevens
David Krech	Eric Trist
Robert W. Leeper	Edward L. Walker
Kenneth MacCorquodale	Robert W. White

The *Psychological Review* is devoted to theoretical articles of significance to any area of psychology. Except for occasional articles solicited by the Editor, manuscripts exceeding twelve printed pages (about 7,500 words) are not accepted. Ordinarily manuscripts which consist primarily of original reports of research should be submitted to other journals.

Because of the large number of manuscripts submitted, there is an inevitable publication lag of several months. Authors may avoid this delay if they are prepared to pay the costs of publishing their own articles; the appearance of articles by other contributors is not thereby delayed.

Tables, footnotes, and references should appear on separate pages; all of these, as well as the text, should be typed double-spaced throughout, in all manuscripts submitted. All manuscripts should be submitted in duplicate. Original figures are prepared for publication; duplicate figures may be photographic or pencil-drawn copies. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail. Manuscripts should be addressed to the Editor, Dr. Theodore M. Newcomb, Doctoral Program in Social Psychology, University of Michigan, Ann Arbor, Michigan.

PUBLISHED BIMONTHLY BY THE
AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
PRINCE AND LEMON STREETS, LANCASTER, PA.
1333 SIXTEENTH ST. N. W., WASHINGTON 6, D. C.

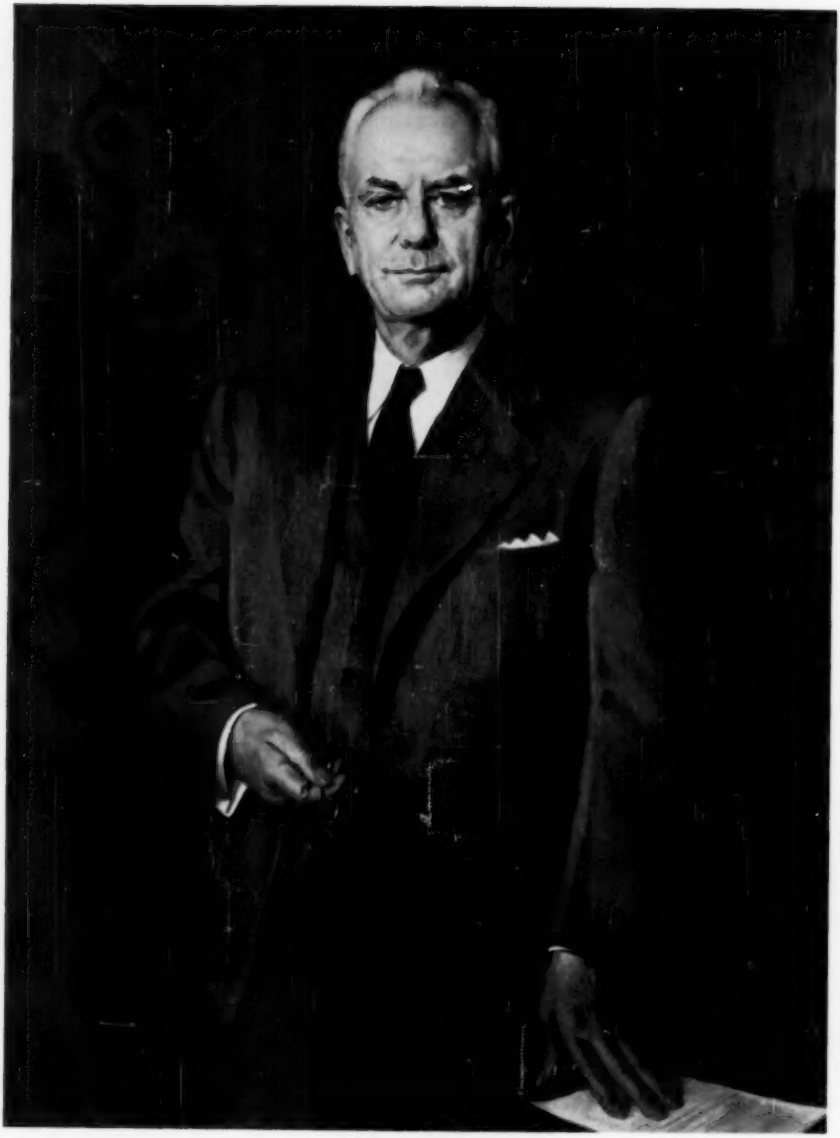
$8.00 volume $1.50 issue

Entered as second-class matter July 13, 1897, at the post-office at Lancaster, Pa., under Act of Congress of March 3, 1879

Acceptance for mailing at the special rate of postage provided for in paragraph (d-2), Section 34.40, P. L. & R. of 1948, authorized Jan. 8, 1948

Send all communications, including address changes, to 1333 Sixteenth St. N.W., Washington 6, D. C. Address changes must arrive by the 15th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

© 1956 by the American Psychological Association, Inc.



WALTER SAMUEL HUNTER

VOL. 63, NO. 4 JULY, 1956

THE PSYCHOLOGICAL REVIEW

WALTER SAMUEL HUNTER
1889–1954

When Walter Samuel Hunter died suddenly on August 3, 1954, American psychology lost one of its most outstanding leaders; Brown University lost one of the most illustrious of its faculty; his colleagues and ex-students lost a delightful friend, and his family lost a fond husband and father.

Walter Hunter was born on March 22, 1889, in Decatur, Illinois. Upon his mother's death when he was 12 years old, his father moved with his two boys to a farm at Saginaw near Fort Worth, Texas. For four years following 1901, Walter attended a two-room school for six months in the winter, worked on the farm in the summer.

Intellectual interests appeared early. During his middle adolescence in Texas, with the sympathetic encouragement of his father, George Hunter, Walter bought and read paper-bound classics of English and American literature, and —probably of more import for his later interests—Darwin's *Origin of Species* and *Descent of Man*. Initiative and leadership also appeared early. At age 15, during the year 1904–1905, he took a correspondence course in electrical engineering, organized in Saginaw an Athenian Literary Society of which he became the first president, and served as local correspondent for the North Fort Worth paper.

It was during a year in the preparatory school of the Polytechnic College at Fort Worth that Hunter was introduced to psychology by his roommate. Reading James's *Psychology, Briefer Course* led to his decision to become a psychologist. After more years at the Polytechnic College, he transferred to the University of Texas in 1908 and began concentration in psychology. Following graduation in 1910 from the University of Texas, he began graduate work at the University of Chicago. The influence of his early reading of Darwin undoubtedly made congenial to him the functional approach of Professors Angell and Carr, and may well have guided his comparative psychological approach to the problem of delayed reaction, which he failed to find in rats but found in increasing degree in raccoons, dogs, and children. His doctoral dissertation, presented first orally in the spring of 1912, established him immediately as an investigator of consequence (1).

Following receipt of his doctorate from Chicago in 1912 after only two years and two summers at the University of Chicago, he returned to the University of Texas as an instructor. Continuing at this double-time rate, he organized his courses which included ethics, Kant, and the history of English philosophy as well as psychology. During January of 1913, before the first semester ended, he married Katherine Pratt of Cleveland and began his research on the aftereffects of visual motion.

Four years at Texas were filled with the teaching of four courses each semester, the direction of master's studies,

and his own investigations. These included studies of aftereffects of visual motion, which related this phenomenon to retinal streaming rather than fading afterimages (2); of auditory discrimination in rats, which resulted in the widely utilized T maze; and of delayed reaction (3, 4). His research during this period resulted in the publication of 10 experimental papers, four theoretical notes, and several reviews. These years were also gladdened by the birth of his daughter, Thayer, and saddened by the death of his wife, Katherine.

In 1916, at the age of 27, he went to the University of Kansas as professor and head of the Department of Psychology. He spent nine years at Kansas. In 1917, he married Alda Grace Barber who had been a former graduate student at Texas, and in March of 1920 their daughter, Helen Barbara, was born. The years at Kansas were interrupted by 16 months of military service in the old Sanitary Corps, spent largely as chief psychological examiner at Camp Lee. There the bar graphs prepared by Hunter and E. S. Jones, showing the predictive value of group psychological tests, influenced a decision in Washington to continue the testing program, and there he influenced the military authorities to order recording of data on the national origins of the troops tested. These data tended to show the relative superiority of North Europeans, but took no account of selective migration. It later irritated Walter Hunter to have these data used in defense of the notion of Nordic superiority and, as he characterized them, "other wild racial theories" (7, 14).

During the Kansas years he produced eight experimental papers which included some of his most important ones. One introduced the temporal maze and the concept of temporal sequences of kinesthetic (feedback, in today's language) stimuli (5). A second introduced the double-alternation procedure and the concept of its dependence upon a symbolic process (9). He also produced a textbook entitled *General Psychology* (1919) and six theoretical papers. One of these (6), even though it then had relatively little influence, appears to be a forerunner of today's concern with the effects of early experience. It pointed out that the stimuli effective in arousing behavior commonly considered to be instinctive can be modified by experience occurring in time before the instinctive organization develops, and suggested that this possibility afforded an important principle for the social control of instinctive behavior. On several occasions during these years, Hunter rejected offers that would have involved him largely in university administration even though these offers were associated with substantial increases in income.

In September of 1925 the Hunter family moved to Worcester, Massachusetts, where Walter had accepted the invitation to become the first G. Stanley Hall Professor of Genetic Psychology at Clark University. During 11 years at Clark he published 21 experimental papers, five theoretical studies, four chapters for books of the Clark University Press, and directed the work of students leading to 22 other publications. By this time he had espoused the behavioristic point of view, but he did not discard the subject's report (10) and he continued to see consciousness (8) and voluntary action (13) as problems for explanation. To avoid popular misunderstanding, he suggested that the term *anthroponomy* replace psychology (11, 12). The amusement aroused by this suggestion he took in good-spirited stride. In this Clark period he also undertook the editorship of the *Psychological Index*, published his second textbook entitled *Human Behavior* (1928), established and began

editing the *Psychological Abstracts* (1927), and served as president of the American Psychological Association (1930–31). He was also elected to the American Academy of Arts and Sciences (1933) and to the National Academy of Sciences (1935).

In 1936 the Hunter family moved to Providence where Walter became professor and chairman of the Department of Psychology at Brown University. There he built a staff known well for both congeniality and productivity, continued as editor of *Psychological Abstracts* till 1946, and served as chairman of the Division of Anthropology and Psychology of the National Research Council (1936–38). This chairmanship led to participation on and the chairmanship of many committees which shortly were activated by concern for the war effort. Hunter served unstintingly, and his influence contributed much to the utilization of psychology by the government and the military services during World War II. For two years (1943–45) he served as chairman of the Applied Psychology Panel of the National Defense Research Council. The citation accompanying his President's Medal for Merit gives Hunter credit for having "recognized that research on the psychological and physiological capacities of man in relation to the new instruments of warfare could contribute materially to the more effective utilization of both military personnel and instruments." He had noted the possibility of such contributions in a letter written to the organizers in 1940 when the NDRC was being formed without any place for the human factor.

Although Hunter continued to give much of his time to committee work for government agencies and foundations, he looked forward to returning to research. On March 22, 1954, his 65th birthday, he had resigned the chairmanship of the department at Brown. While on a trip to visit the families of their daughters, the Hunters stopped at the writer's home about three weeks before Walter's death. At breakfast, Walter described his release from administrative and committee activities and welcomed his return to research and teaching with, "I'm going to enjoy these next five years," and then with an impish grin, he added, "maybe ten, maybe twenty."

In each of his various professional roles as investigator, as theorist, as teacher, as editor, as servant of his professional associations, as policy-making committeeman for government and foundations, Walter Hunter made substantial contributions. Yet these contributions are not the full measure of his professional life. Many a student and younger colleague owes both his scientific conscience and his professional confidence to Hunter, who could be critical of an argument or plan of attack, sometimes devastatingly so, and yet encourage the person, be he student or staff member. In presenting an assistantship or job to a prospect, he played up the difficulties and played down the advantages, with the result that things regularly worked out better than student or staff member had been allowed to expect. Hunter delegated easily. Once assignments were agreed upon, he supervised only when asked to or when trouble had obviously developed. He took pleasure in the accomplishments of his students and staff members, as he also did in his own, and once remarked that the chief satisfaction of an academic administrator was to "make great men of his boys." This pattern was apparently one of long standing, for one of his staff members from the Kansas days has remarked that Hunter re-established her faith in people, a faith that had been shaken in

a devastating experience of graduate work. With this mode of operation, no special school of thought developed around Hunter. Each member of his group remained free to solve his own problems and develop his own theories. Yet Hunter's leadership brought about that rare combination of congeniality and prodictivity for which the Brown group of the last 20 years has been known.

Walter Hunter's life offers a number of intriguing paradoxes. Even though his rate of productivity was exceedingly high, he seemed ever to be relaxed and to have time to spare. Although he avoided administration, he was highly successful at it, as shown by the success of both the Brown Department and the Applied Psychology Panel. Although he tended verbally to emphasize the impersonal and the objective, he was warm in his own interpersonal relationships and wise in leadership. Although he accepted the behavioristic point of view, at least methodologically, and espoused it strongly, one of his important investigative contributions was the demonstration via double alternation of a symbolic process in animals. Even though he was an experimental psychologist who often minimized the importance of professional psychology, he devoted the last decade of his life to the Applied Psychology Panel and to committee work which resulted in a tremendous growth in professional psychology. Perhaps these paradoxes derive from his concern for the problem and for the empirical. He respected theory as a set of constructs subject to revision, but he distrusted doctrine.

For all Walter Hunter's productivity, his professional work never absorbed him completely. He took up painting and developed skill at it. This art he shared with his wife, Alda, in whose surpassing skill he took great pleasure. During the early Brown years, he and Mrs. Hunter joined with younger colleagues and others in a dancing class and club. He enjoyed Ping-pong, and he competed evenly with the best of his younger colleagues. The Hunter house, built in 1938, had a playroom devoted to Ping-pong. A party at the Hunters' during those days included good food, good talk, and usually Ping-pong. These parties were fun.

Walter Hunter's frank enjoyment of an honor was a delight to behold. It is gratifying to his friends that he had seen the painting done for his retirement (a print of which appears herewith), and enjoyed it thoroughly. He would have enjoyed the ceremony connected with its unveiling at the meeting of the Society of Experimental Psychologists in the spring of 1955. He would also have enjoyed planning the new building for the Psychology Department at Brown, a still unfulfilled hope when he died, but he had lived a good, full, and highly useful life. On the shoulders of such giants do we build the future.

REFERENCES

1. Delayed reaction in animals and children. *Behav. Monogr.*, 1913, **2**, 329–333.
2. The after-effect of visual motion. *Psychol. Rev.*, 1914, **21**, 245–277.
3. Auditory sensitivity of the white rat. *J. anim. Behav.*, 1914, **4**, 215–222.
4. Auditory sensitivity of the white rat. *J. anim. Behav.*, 1915, **5**, 312–329.
5. The temporal maze and kinaesthetic sensory processes in the white rat. *Psychobiol.*, 1918, **2**, 339–351.
6. Modification of instinct from the standpoint of social psychology. *Psychol. Rev.*, 1920, **27**, 247–269.
7. Relation of degree of Indian blood to score on the Otis Intelligence Test. *J. comp. Psychol.*, 1922, **2**, 257–277.
8. The problem of consciousness. *Psychol. Rev.*, 1924, **31**, 1–31.
9. The symbolic process. *Psychol. Rev.*, 1924, **31**, 478–497.

10. The subject's report. *Psychol. Rev.*, 1925, **32**, 153–170.
11. General anthroponomy and its systematic problems. *Amer. J. Psychol.*, 1925, **36**, 286–302.
12. Psychology and anthroponomy. In *Psychologies of 1925*. Worcester, Mass.: Clark Univer. Press, 1926. Pp. 83–107. *Ped. Sem.*, 1926, **33**, 322–346.
13. Voluntary activity from the standpoint of behaviorism. *J. genet. Psychol.*, 1934, **10**, 198–204.
14. Walter S. Hunter. In *A history of psychology in autobiography*. Vol. 4. Worcester, Mass.: Clark Univer. Press, 1952. Pp. 163–187.

J. McV. HUNT
University of Illinois

Psychological Review
Vol. 63, No. 4, 1956

ON THOUGHT: THE EXTRINSIC THEORY [1, 2]

EUGENE GALANTER [3] AND MURRAY GERSTENHABER

University of Pennsylvania

Introduction

To say that an organism "thinks" is to say that it changes its behavior in anticipation of a changing environment. The *behavior* occurs, of course, only if the organism is motivated to display its prediction about the environment. Some say thinking may occur with no behavioral accompaniment, or with at most a verbal report of introspections. While this kind of "thinking" may be very important in the control of behávior at some later time, an objective theory of thinking must describe behavior concomitant with the thought process. One essential ingredient of a behavioral theory of thought, therefore, is an analysis of the role of "motivation." Although many psychologists have pointed out the need (4), the role of motivation in thinking has not been systematically clarified. The primary element which any operational model of a theory of thinking must include is a motivating technique to make overt the covert processes.

Before we go on to the development of such a model, a classification of the current formulations of the problem of thought is useful. Classical theories of thinking fall into two general (and oversimplified) categories. These classes of theories have been termed variously (*a*) S-R, association, trial-and-error, or sign theories; and (*b*) image or model theories. On the whole, S-R theories account adequately for changes in behavior when the environment remains constant while time passes or trials take place, i.e., "learning" occurs. Sign theories also account for failures of thought. For example, S-R associations to a specific S continue to produce R when perceptually irrelevant, but eventually important, components of S change. The emitting of R, although now inappropriate, is usually laid to "habit." However, these trial-and-error theories give a poor account of complex problem solving; in matters of this kind, trial-and-error theories are either inelegant or equivocal.

These criticisms of S-R theory should not be taken to imply the acceptance of image theories, for images are not generally more successful as an explication of thoughtful behavior. It is difficult to specify the behavioral manifestation or identification of an image, and in addition, image theories rarely give more than a slipshod account of how the image comes into its full-blown existence. Even so, the fuzzy description of the role of images in complex thinking does a fuzzy kind of justice to our introspections.

In order to avoid the pitfalls of sub-

[1] Portions of this paper were read at the American Psychological Association meetings in San Francisco, 1955.

[2] The authors are indebted to Dr. Elliott Danzig for his generous cooperation in the early stages of the research, and for his illuminating and suggestive ideas concerning the thought processes, many of which have found their way into the present manuscript. The Office of Naval Research made available to us a transcription of comments on parts of the material included in this paper when it was presented to their research group in the Spring of 1955. The pilot research mentioned in the present paper was supported by a grant from the Faculty Research Fund of the Graduate School of Arts and Sciences of the University of Pennsylvania.

[3] Research Fellow, Harvard University, 1955–56.

scribing to one or the other of these two kinds of theories exclusively, most early theorists attempted a synthesis. Such synthesizing went out of vogue during the "behavioral revolution," and now most psychologists espouse either an S-R theory or a cognitive theory. Our ultimate aim is to try again the development of a synthetic theory. This synthesis must contain two components, an S-R theory and an image theory. Since most of the current model building has dealt with S-R concepts, and since a well-developed and workable model of behavior of this kind exists (1, 2), our first concern is with the development of an image theory of thinking. The main problems of this paper will be (*a*) to describe a motivational technique for making the latent image behaviorally manifest, and (*b*) to state an hypothesis about the relation between S-R and image theories.

In order to help the reader to follow what may otherwise seem a clumsy development of the theory, we shall first present an intuitive notion of imaginal thinking. Imaginal thinking is neither more nor less than constructing an image or model of the environment, running the model faster than the environment, and predicting that the environment will behave as the model does. To refine this concept we will need to define certain common terms. The phrase "required to think" will be used of a motivated organism whenever it faces an environment that is able to yield some reward or value. Presumably the organism does not know the particular *sequence* of actions required in order to achieve the reward. The environment will be called a "machine" or "mechanism." Thus we say that an organism is faced with a problem if it is motivated to *predict the behavior* of the machine. The process by which the behavior of the mechanism is predicted is called "thinking," and adequate (in the sense of achieving the reward) prediction is called a "solution." When the sequence of acts is unitary (e.g., the Skinner box), the solution is often accomplished by trial and error, but when the sequence is more complicated, the solution usually involves the use of imagery. Note, however, that even a unitary sequence is a sequence, and therefore in this degenerate case an "S-R connection" is an image, and/or vice versa.

The first step in the solution of a problem consists in the construction of a model or image of the "relevant" features of the machine. These models can be constructed from many things, including parts of the organic tissue of the body, and by man, paper and pencil, or actual artifacts. Once a model has been constructed, it can be manipulated under various hypothetical conditions and constraints. The organism is then able to "observe" the outcome of these manipulations, and to project them onto the environment. If the image mirrors the relevant aspects of the environment, prediction is possible.

According to this view, an aeronautical engineer is thinking when he manipulates a model of a new airplane in a wind tunnel. The motorist is thinking when he runs his finger over a line on a map, the finger serving as a model of the relevant aspects of the automobile, the map as a model of the road. External models of this kind are often used in thinking about complex physical machines. Images used in covert thinking depend on the availability of the physico-chemical elements of the organism which must be used to form models.

The task is to discover how the model-building enterprise proceeds and is controlled. In order to describe the construction of the model of a problem, it is necessary first to specify what constitutes a "problem," and then to show

how experiments that satisfy these specifications can be constructed and performed. As a first step we can say that a problem exists when an image of the environment exists, but this is clearly circular. Before we can use the existence of an image as a criterion for the presence of a problem, we must agree upon a behavioral criterion for the "existence of an image." This concept of "image existence" is analogous to, if not identical with, the classical notion of "insight." In fact, the identification is so close that we can proceed directly to examine the idea of an insight. The analysis of insight is a prerequisite to the specification of a "problem," and is also necessary for understanding the complementary relation between the image and the S-R theories of thinking suggested above.

Since the time of Köhler's classic work (6), it has generally been accepted that a major behavioral characteristic of insight is a radical change in the behavior involved in the achievement of a goal.[4] But this criterion is not sufficient as a general operational definition of insight. In the first place, radical change in behavior does not necessarily mark insight; for example, a person attempting a "bent-nail" puzzle may, after a number of wholly random trials, separate the links. If he has noted his motions, he may thereafter solve the problem immediately. The radical change observed in his behavior does not necessarily imply insight, nor is insight precluded. It is equally true that insight is not necessarily accompanied by a radical change in behavior. Sometimes a problem solver eliminates an obviously absurd hypothesis from among countless alternatives. This may be insightful, but it will produce no *detectable* change in behavior. Finally, the behavior that follows upon insight may not be *obviously* goal-oriented—the goal the subject sees may differ from that set by the experimenter.

The difficulties encountered with our initial characterization of insight suggest three criteria for problems designed to enable us to examine thinking. The first can be called the "stimulus" criterion, and it states that the problems must require an insight that is difficult of achievement to a prescribed degree. In practice there must be a method for creating problems of any degree of difficulty or complexity, and a measure of the difficulty or complexity which is not circular. The second criterion is the "response" criterion. Any insight must be reflected behaviorally within the problem situation. That is to say, the performance must be quantitatively measurable. This means that the number of alternatives presented to the subject must be severely restricted. Most situations cannot satisfy both criteria simultaneously because complexity is normally achieved by increasing the number of alternatives. The third criterion will be called the "motivation" criterion. The way in which insight will manifest itself must be known in advance. In particular the transfer or stimulus-generalization properties of the image must be measurable for a variety of situations. When the third criterion is satisfied, it becomes unnecessary to ask the subject for a verbal report in order to determine his goals.

Of the experimental techniques used in the study of problem solving by human beings, the type that best meets these three criteria involves binary choices. These techniques satisfy both stimulus and response criteria, for complex sequences of binary choices can be manufactured at will, and the performance of the subject can be quantitatively speci-

[4] We wish to avoid discussion of the other behavioral characteristics of insight which Köhler utilized in his construction of the most trenchant form of the image theory in the present century.

fied. Only the third criterion appears not to be met by the use of binary choices. We will not consider the needed improvement of the technique at this point, but will turn rather to a short description of the standard form of the binary-choice experiment.

Methods that use binary choices provide for the repeated presentation of an alternative to which the subject can make one of two mutually exclusive responses. After each response, the subject discovers whether the response was correct. With such a procedure two outcomes are possible: the subject who makes a wrong response does not find out whether the other would have been right; or, regardless of his response, he is told which event occurred. We shall restrict the present discussion to the second alternative: after he has made his choice, the subject is always told which response is correct.

A machine may be programmed to generate various kinds of binary sequences. In their intuitive order of complexity these include constant-valued sequences (all zeros correct, or all ones), the alternating sequence 010101 . . . , the double alternation 001100110011 . . . , etc. To any of these sequences various kinds and amounts of "noise" can be added to produce even more complexity. Those sequences are said to be noisy in which the next element in the sequence is determined by a probability mechanism. The generality of this experimental paradigm resides in the fact that any problem expressible in "natural" language can be expressed by such a sequence. But a more important reason for selecting the two-choice situation is its simplicity and convenience.

Experiments that utilize two-choice situations have generally violated all the criteria mentioned above. Most of these experiments, however, were designed with different aims. The use of problems based on binary choices does not add to our understanding of the thought process, unless the problem is so constructed that the "stimulus" is measurable, the "response" is unequivocal, and the "motivation" is specified, i.e., the problem meets the three criteria.

THE PAYOFF FUNCTION AND THE DEFINITION OF A PROBLEM

When a person faces a problem, his behavior is determined by his expectation of reward. "Reward" should be construed to mean the net value that accrues from a course of action. Some of this value may consist in an extrinsic reward (food or money); other components may include the intrinsic tastiness or difficulty of a particular act. A further advantage of our concentrating on binary choice is that the dependence of the value of a reward on response preferences or aversions is minimized (though not eliminated).

We may assume that we have defined a function whose domain of definition is the set of all possible activities of the individual, and which assigns to each element in the domain the reward to the individual for adopting that activity. This function will be called the *payoff function* (Fig. 1). In most psychologi-

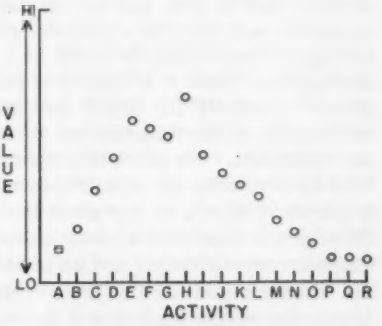


FIG. 1. Payoff function (general).

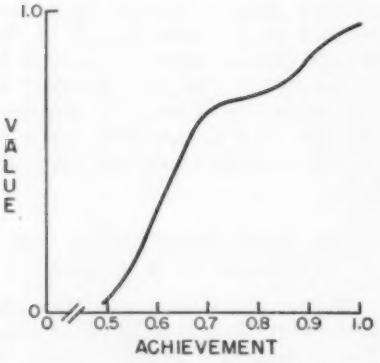


FIG. 2. Payoff function (specific).

cal experiments this function is unknown. An experiment will here be called a *problem* if the subject can determine the range of values of the payoff function and the reward for a particular choice of activity, after he has made his choice. For simplicity it will be assumed that a real numerical value between zero and one can be assigned to the reward. Two choices will be considered equivalent if the rewards they produce are identical. Thus the activities can be ordered so that the graph of the function never decreases.

When the alternatives are binary, the subject's decision is expressible as a sequence of binary choices. In this situation the payoff is dependent on the subject's choices only, and not on any extraneous activity. In a problem involving n binary trials, there will be 2^n alternatives. When n is large, it is not possible to specify the payoff function unless many of these alternatives are in fact equivalent. We effect this equivalence by restricting our attention to experiments in which, on any given trial, the subject is faced with a binary choice with one correct element; and we postulate that the payoff function takes as its domain the relative frequency of correct choices. For a large n, this is in some sense the asymptotic percentage of correct choices (which we assume exists). On the assumption that n is large, the payoff function can be represented by a graph on which reward is plotted against asymptotic percentage of correct choices. This percentage will be called the subject's *achievement* (Fig. 2).

The notion of achievement is definable for a class of problems larger than sequential binary-choice problems, but not for all classes. What follows is meaningful for all problems where the notion of achievement is definable.

CHARACTERISTIC LEVELS OF INSIGHT: THE ACHIEVEMENT DISTRIBUTION INTEGRAL

It is now possible to ask whether the criteria stated earlier are satisfied by the problems of a given class. The first criterion is crucial for the detectability of insight, but it depends on the notion, undefined as yet, of "characteristic level of insight." We will say that a problem *characteristically requires at least some level of insight* if different subjects attempting the problem differ in achievement. That the problem requires insight is shown by the variation of achievement. Since achievement is an asymptotic value, we may expect the distribution of subjects' achievements

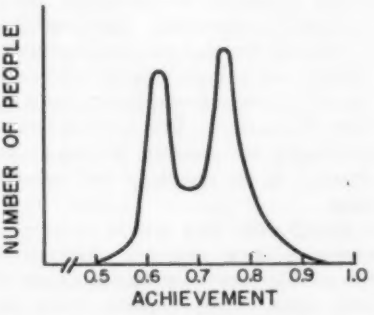


FIG. 3. Achievement distribution curve.

to be multimodal with a narrow distribution about each mode (Fig. 3). The problem may be considered to have been solved at different levels by the persons distributed around each of the modes. To the solution of the problem at each of these levels may be assigned a characteristic level of insight, namely, the percentage of people who do not solve the problem at any higher level.

All actual, sequential, binary-choice experiments must terminate after a possibly large but nevertheless finite number of trials, and the approximation to the achievement determined by this finite number can be expected to show a positive variance even when none would be present in the limit. It is reasonable to assert, however, that a significantly multimodal distribution of persons over approximate achievements indicates that the problem requires insight. The graph of the distribution of persons over achievement will be called the *achievement distribution curve* of the problem, and its integral, normalized to have a maximum value of one, will be called the *achievement distribution integral* (Fig. 4). Insight is manifested in the achievement distribution integral by the presence of plateaus.

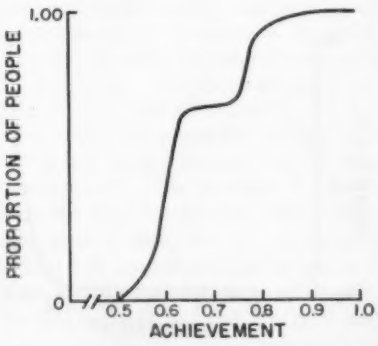


FIG. 4. Achievement distribution integral.

DEPENDENCE OF THE ACHIEVEMENT DISTRIBUTION INTEGRAL ON THE PAYOFF FUNCTION

We have stated that a problem consists of an external situation that allows a subject a certain number of choices (called a machine), and a payoff function. In the problems we consider, the payoff is a function of the objectively determinable achievement, and this function is, by the definition of a problem, known to the subject. The classical notion of a problem did not explicitly involve a payoff function, and experimenters sought to determine directly from the reactions of a subject faced by a machine whether insight was present. The insight sought concerned the nature of the machine, and this is what most investigators of thinking ultimately attempt to isolate. But whether explicit or not, the payoff function is always present, being communicated by instructions or lack of them. It is necessary, therefore, to investigate, for a *fixed* machine, the dependence of the achievement-distribution integral on the payoff function. That is to say, we must determine the effects of different instructions on the presence or absence of insight, when the environmental situation is constant. This dependence is beautifully demonstrated in an experiment by Goodnow (3), and it is on these and similar data that the following discussion is based.

The dependence of the achievement-distribution integral on the payoff function is a fortiori a function of the particular machine being used. For a fixed machine, however, this dependence is represented by the entire behavioral manifestation of all the insights of the experimental population. Specifically, those insights that are characterized by this dependence must concern the nature of the machine. The second criterion for problems that test insight—

i.e., that any insight must be reflected in behavior—therefore holds for the class of problems with a common machine, provided the dependence of the integral on the payoff function is known. At present, however, the dependence is known only after the fact, and the real question is whether it can be predicted in advance. To make such a prediction, a fundamental hypothesis must be made about the nature of the individuals through whom this relationship exists. This hypothesis needs quantitative refinement, but it may for the present be stated in the following qualitative fashion:

Hypothesis: The likelihood that an individual will construct an image of a machine increases when the cost of S-R or sign behavior increases.

Some examples of the experimental implications of this hypothesis will best reveal its general import. Consider first the individual faced with a binary-sequence generator producing zeros and ones according to the principle that a one will be produced at random 75 per cent of the time (i.e., a constant-valued sequence with 50 per cent noise). Suppose further that there is no payoff function, but the subject receives one penny for correct responses and nothing for incorrect ones. We know from previous research (5) that the subject will tend to match his response frequency to the event frequency. After the subject has reached stability, say at 100 trials, we select 10 per cent of the trials that follow, making our selection at random, and for these trials tell him that for correct responses he will receive ten cents rather than one. His performance prior to this time is characterized by S-R learning; in fact it is described quite well by a sign-learning theory like that of Estes (2) or of Bush and Mosteller (1). For 10 per cent of the remaining trials we have now increased the cost of this behavior, and we predict that, at least on these trials, his behavior will change, i.e., he will select the more favorable event (the "one") more often than it is likely to occur. Unreported experiments by Professor Sidney Siegal, of the Pennsylvania State University, and by Galanter support this conjecture. In addition, in Galanter's experiments, the subjects tended to increase their selection of the more probable event throughout the remaining trials. It is an interesting parametric question how much of an increase in reward will produce insight for all trials.

Alternative methods can be used to demonstrate the same effect. Thus when stability has been reached, the subject can be told that after a particular selection he must respond in the same way for one more or two more or n more trials. Again, the cost of trial and error has been raised, and again we predict insightful behavior. As a final example of this type, the subject can be negatively reinforced for switching responses after he has reached stability.

As a more powerful example of the implications of the hypothesis, and as a demonstration of the formal dependence of the achievement-distribution integral on the payoff function, consider an experiment reported by Smith and

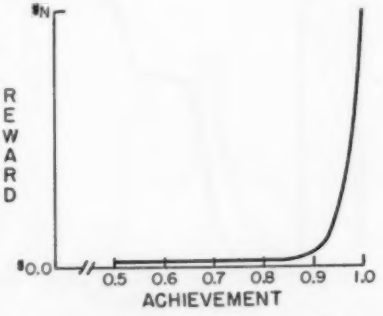


FIG. 5. Payoff graph—problem "P."

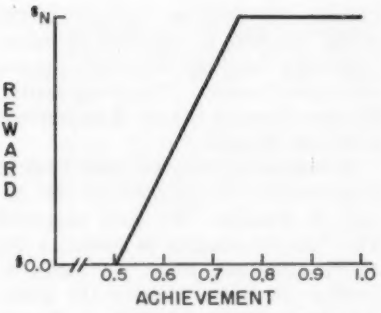


FIG. 6. Payoff graph—problem "Q."

Galanter (8). A machine generates zeros and ones as before. With this machine two problems are defined, P and Q, with the following two payoff functions. The value of P equals zero when the achievement is less than or equal to 90 per cent, and the value of P is maximal when the achievement is greater than 90 per cent (Fig. 5). For problem Q the value is four-thirds of the achievement when the achievement is less than three-quarters, and the value of Q is maximal for an achievement greater than three-quarters.

We assume, of course, that value is linear in money, even though this assumption may be counter to the facts. But for the simple experiments considered here only monotony is required. With more complicated problems it may become necessary to measure the value of money to each subject by some scaling technique. In the previous problems the random elements in the machine made it impossible for the subject to get an achievement higher than three-quarters. In problem P above, since the slope everywhere in the achievable range is zero, the cost to the subject of S-R experimentation is low, and from the hypothesis one predicts that the subject will proceed by trial and error. By the nature of the machine this must result in a mean achievement smaller than three-quarters, and this is borne out by the experimental evidence. For 15 subjects who were given problem P the achievement averaged 63.4 per cent. In problem Q, because of the large slope in the achievable range, the cost of experimentation is high, and the subject, finding that he can achieve three-quarters, will in fact behave to obtain it. For these 15 subjects the mean achievement was 72.3 per cent.

RATIONAL PAYOFF AND NORMALIZED PROBLEMS

In problem P of the preceding section, one would not expect any characteristic level of insight to be manifested (the distribution integral shows no plateaus). The fact that this has been observed experimentally has usually been considered an anomaly, exhibiting as it does seemingly irrational behavior. It is not. The "irrationality" is this: Many subjects upon interrogation are found to know that an achievement of three-quarters is obtainable, and they know how to obtain it, but the experimental results do not differentiate between those subjects who presumably have some insight into the nature of the machine, and those who do not. The irrationality lies actually in the payoff function. The payoff for a given achievement is in no way commensurate with the difficulty of the achievement, and the result is that no information about the subject's insight is gained. We learn a little about the dependence of the distribution integral on the payoff function, but our net conclusion is that this problem is not fruitful for the study of insight.

Insight may manifest itself in a personal interview but not in the behavior in the problem, and this violates the third criterion. The answer to the question of how to find a suitable payoff function for a given machine would

seem to be to let the payoff for a given achievement be the difficulty of obtaining it. This in turn is measurable only by the percentage of people obtaining that achievement, which in turn depends on the payoff function, and the whole argument turns in a circle! What is desired, however, is only that the graph of the payoff function for a given problem be identical with its distribution integral. Such a problem will be called *normalized*. For normalized problems the payoff is "rational," the payoff for a given achievement being equal to the difficulty of obtaining it, as estimated by the proportion of the population that has achieved at this level. It seems reasonable to expect that problems like this have maximum likelihood of exhibiting characteristic levels of insight. It is our conjecture that it will be found characteristic of normalized problems that their distribution integrals exhibit plateaus, and it is our thesis that, at least in the simple type of problem considered here, the existence of plateaus is the criterion for the manifestation of insight, and tells where to look for such manifestations. It would be interesting if normalized problems could be produced experimentally. It is likely that they can be obtained by a method of successive approximation.

Now, given a fixed machine, the question of the dependence of the distribution integral on the payoff function can be reduced to the simple question: How can we construct normalized problems? The question is not easily answered, for it is conceivable (and in fact likely) that two distinct payoff functions coupled with the same machine may both produce normalized problems. What we are looking for is a model for describing "problem complexity." This notion is at the heart of the matter, for a model of this kind will be forced to make assertions about the structure of hypothetical mechanisms that examine sequences, and then, on the basis of the payoff function, to construct sequence-generating functions to predict the environmental events. The whole matter, however, is properly part of an intrinsic theory of thought.

In summary, then, we have tried to conceptualize the problem of the nature of thought. We have suggested that thought consists in modeling the environment and using the model to predict the future state of the world. To examine the mechanism of thought we must force the organism to display this function behaviorally. By introducing the idea of a payoff function, we can control the motivational input to the organism, and so increase the likelihood of having its insights manifested in action. We then proposed an hypothesis relating S-R behavior and insightful behavior, and described some scraps of evidence in its support. Finally we pointed to the need for an independent measure of "problem complexity" before we could get to the critical problem of the relation between the motivational structure of the person and the characteristics of the insight he reveals.

In conclusion, it should be pointed out that the program stated in the beginning of the present paper has not been completed. We have concerned ourselves only with identifying the existence of images, and have not even suggested how they develop. The transient states (the "learning") have been neglected. This problem is resolved in a natural way when the mechanism for describing "complexity" has been constructed.

REFERENCES

1. BUSH, R. R., & MOSTELLER, F. *Stochastic models for learning*. New York: Wiley, 1955.
2. ESTES, W. Toward a statistical theory of learning. *Psychol. Rev.*, 1950, **57**, 94–107.

3. GOODNOW, J. J. Determinants of choice-distribution in two-choice situations. *Amer. J. Psychol.*, 1955, **68**, 106–116.
4. HUMPHREY, G. *Thinking*. New York: Wiley, 1951.
5. HUMPHREYS, L. G. Acquisition and extinction of verbal expectations in a situation analogous to conditioning. *J. exp. Psychol.*, 1939, **25**, 298.
6. KÖHLER, W. *The mentality of apes*. New York: Harcourt, Brace, 1925.
7. PRICE, H. H. *Thinking and experience*. Cambridge: Harvard Univer. Press, 1953.
8. SMITH, W. A. S., & GALANTER, E. H. A test of a new theory of thinking. *Amer. Psychologist*, 1955, **10**, 487. (Abstract)

(Received December 16, 1955)

Psychological Review
Vol. 63, No. 4, 1956

A NEW THEORY OF COLOR VISION

WILLIAM A. SHAW

University of Pennsylvania

In 1928 Hecht (11) argued that Young's hypothesis concerning the presence in the eye of three kinds of color receptors could be "a basis on which to build a theory for the mechanism of color vision" if the "uniqueness of the sensations of yellow and of white" (p. 238) could be accounted for. He set up a simple experiment to answer the question, "Are there special substances in the retina for the reception of yellow and white, or are yellow and white phenomena which arise in the brain out of the impulses coming from the three kinds of fibers (or substances) in the retina?" (p. 238). He proposed to show by stimulation of corresponding points of the two eyes with red and green lights, respectively, that "the red sensitive fibers and the green sensitive fibers are both active, and the brain synthesizes yellow, as must be the case if the Young-Helmholtz idea is adopted" (p. 238).

Hecht's experiment confirmed his belief in the existence of binocular yellow. His success in performing this critical experiment is overshadowed by the many failures in the past, and by the severe criticism of both his apparatus and his method by later experimenters.

The present paper is not so much interested in the outcome of binocular color mixing as it is in pointing out how narrowly color theorists tend to remain within the framework of classical theory. The great Helmholtz was ignorant of the many facts at Hecht's disposal. Hecht did not seem to be aware that his speculations preclude the existence of the yellow-blue blind. He appears to have been extraordinarily shortsighted; first, in selecting only a single item from among the many facts at hand to work on, and secondly, in believing that an account of binocular yellow could render the Young-Helmholtz theory workable. Had not Ladd-Franklin (21) annihilated Helmholtz years before? Helmholtz's theory does not suffer merely because of its difficulty in accounting for the "uniqueness of yellow and of white" but because of the assumption that biological irritability results in sensations. There should be no more mystery about the yellow of a red-green mixture than there is about the blue of a green-violet mixture.

Other theorists, although dissatisfied with the Helmholtz theory, have not been able to free themselves from it. Hartridge's (10) polychromatic theory illustrates this dissatisfaction and at the same time demonstrates how forcefully classical theory grips any attempted new formulation. Hartridge believes that the Helmholtz theory suffers because the initial investigators helped to fix the number of cone types at three. He maintains this position in spite of the fact that the results of his microstimulation studies *can* be interpreted, not as evidence for more cone types, but as evidence for no cone types at all. Similarly the excellent microelectrode studies of Granit (7) can as easily be interpreted to point away from as toward the acceptance of special color cones.

The Young-Helmholtz, the Ladd-Franklin, and the Hartridge theories all have a common "central" feature. The mixed hues, following upon a special retinal alchemy (21, p. 457), are somehow manufactured in the brain. All are concerned with fixing the num-

ber of cone types. The theories mentioned here are not unusual; indeed, they are typical. Göthlin (16), for example, assembles a great amount of evidence against the possibility of a yellow receptor. He concludes that yellow is synthesized from red and green, but fails to explain why, in the tritanope with usable red and green sensitivity, the yellow region is reported as colorless.

In many ways a greater freedom for investigation and interpretation is shown in the studies of the rod substances by Hecht (12, 13) and by Wald (31, 32). This trend ought to be emphasized in the cone studies. Unfortunately the psychologist is rarely equipped to study receptor substances directly. He is usually limited to the study of the behavioral effects of stimulation. Although desperately in need of information of other kinds, his case is not hopeless. If nothing else, he can act as a gadfly, a capacity in which Ladd-Franklin (20) served superbly.

On the behavioral level, then, it appears to the author that there are approximately four groups of data of the greatest importance to color theory. In retrospect, the reasons for selecting these particular data are distressingly vague, and show little more than the author's desire to emphasize the fact that in the data of color mixture there are relationships which have been obscured and ignored by forcing the idea of cone types upon empirical data. They are: (*a*) the fact that in a mixture of complementaries the hue that is seen is one or the other of the complements, depending on their intensities; (*b*) the fact that the spectrum is divided in the green area in such manner that on either side of this range of frequencies a mixture of any two frequencies will produce an intermediate hue (1), but that a mixture of any two frequencies, one each from the two halves, will not yield an intermediate hue; (*c*) the existence of complementary frequencies and their positions in the visible spectrum; and (*d*) the fact that the saturation of the mixture of two or more frequencies is nearly always less than that of the spectral hue which it matches.

Still a fifth possibility was considered, namely, that any spectral hue can be matched by the proper selection and mixture of three primary hues. In spite of the excellent work of Wright (35), Guild (8), Hardy (9), and Judd (15), who provide us with the means of relating and comparing the work of different investigators, the common form of the above statement leads to its rejection. This rejection is based upon the distinction between real events and imaginary ones. The real events concern the manipulation of physical light mixtures.

What are the facts? Three real primaries cannot be mixed in any proportions greater than zero amounts to match a standard. In a sense, the operations performed define the significance of the results for color theory. The transformations which have been imposed upon the real primaries, and which change them into imaginary ones, deprive the real primaries of their significance. The tristimulus values of the various sets of primaries have been adjusted to fit the needs of the various experimenters. The resultant curves are not descriptive of the underlying physiological processes. What possible meaning can be extracted from these curves relevant to an explanation of "white" (19)?

How the mixture of complementaries depends upon their intensities. In regard to those pairs, which are complementary, it is commonly noticed that they do not add together like any of the other hues. For example, for the yellow-blue pair, if one adds blue to the initial yellow, the mixture acts as

though one subtracted a certain amount of yellow, i.e., desaturated it. A point is eventually reached when the intensity of the two hues has the same ratio as do their chromatic thresholds (25), at which time the mixture will appear colorless. Adding more blue at this time makes the mixture look bluish, with no trace of the initial yellow.

To say that the complements stimulate their respective cones equally is not an explanation for the disappearance of the hue. It is not a universal explanation for mixtures, since non-complementary mixtures resulting from equal stimulation do not show this effect. From this we infer that, since the complements are always selected from the opposite sides of the green in the spectrum, they originate some processes which are in a sense opposed.

The spectrum is divided in the green. Concerning the second item of the initial observations, nothing can be learned from the transformed data on color mixing. The hypothetical red, green, and blue curves admittedly do not describe any real cone responses or processes. The actual facts are that, throughout the spectrum, narrow bands of frequencies lying relatively close together can be mixed to mediate hues which cannot be discriminated from hues intermediate between the two. Again, frequencies on one side of green —for example, the longer wavelength side—irrespective of their proximity, can be combined two at a time to produce an intermediate hue. The middle range of frequencies, green, is common to both halves, mixing with red to produce a yellow, and with a violet to produce a blue. The saturation of these mixtures is almost always less than that of the spectral hues. Thirdly, one cannot mix frequencies taken one each from the opposite sides of green and secure intermediate hues. Certain of these pairs form the well-known complementary hues. Others, however, when mixed will give a hue lying on one side or the other of the spectrum but not between the two members of the pair (18).

The significance of this, I believe, is the support it lends to the idea that the hues on the opposite sides of the spectrum are somehow opposed to each other. As long as we stay on one side or the other of the spectrum the frequencies have some property which is alike but opposite to the other side.

Complementary frequencies and their positions in the spectrum. Priest (24) summarized the various sets of complements and plotted the pairs against each other. He developed the following empirical formula (530–f1)—(f2–608) equals 220. Later Grunberg, according to Southall (28, p. 317), proposed a formula based on wavelength; (wavelength-559)—(498-wavelength 2) equals 424. Both can be used to estimate the complement of a given frequency or wavelength.

The regularity with which the distance between the members which lie on the long wavelength side of the region of maximum sensitivity decreases, compared to the increase in distance between the other members which lie on the short wavelength side, is more striking if one marks the positions of these complementary pairs on a scale of wavelengths as illustrated in Fig. 1. It is with this observation that we begin our inquiry. This orderly array, which could not have occurred by chance, could not be without meaning. That the pairs form an equilateral hyperbola may have no significance

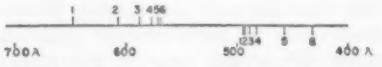


FIG. 1. The pairs of complements are located on a scale of wavelengths in order to show that as one approaches the green region from one side, the complement recedes toward the limit of visibility on the other side.

from the standpoint of correlation with physiological events. On the other hand, if such a correlation exists, it might be obscured by this form of presentation. Tentatively, then, rather than neglect this observation, as present theories do, let us assume that there is some physiological correlate connected with it, and explore it more fully.

The relationship of the complementary pairs is sometimes shown by plotting wavelength against wavelength (34, p. 552). The resultant curve is not very informative. The step not taken by the psychologist is one relating wavelength to velocity. Wavelength as we commonly speak of it is simply a function of frequency with velocity held constant. The psychologist never questions this. As a rule he knows only one added bit of information, namely, that the reason a glass prism works is because the frequencies do not travel through it at the same rate of speed. Regardless of the relation of the wavelengths of the complementary pairs in air, there is certainly every possibility that these wavelengths are different as the various frequencies pass through the absorbing substances of the retina. In other words, the frequencies possess a property, namely, velocity, which varies according to the media through which the frequencies travel.

The mere idea that the visible frequencies have different velocities when traveling through such substances as water or glass strikes no sparks. The substance in the retina, however, is like neither of these. It is absorbing. How is velocity altered by an absorbing substance? For the aniline dyes the velocity of the frequencies on the long wavelength side of the spectrum is remarkably decreased, whereas for those on the short wavelength side it is just as strongly increased as one approaches the absorption band from the ends of the spectrum (22, pp. 118–120; 33, pp. 113f.). From this it follows that a spectrum from a prism filled with such a dye would not show the same arrangement of the hues as that resulting from passing the light through a glass prism. The dispersion is markedly anomalous, as shown in Fig. 2.

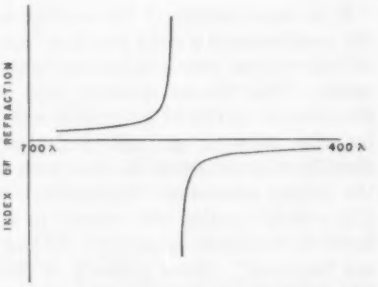


FIG. 2. A curve of refraction of a substance in which the absorption band falls near the middle of the range of visible frequencies.

Now the substances in the retina are not only absorptive; they are also optically active. (Wood [33, p. 300] notes that the fovea is differentially sensitive to the ordinary and extraordinary rays of polarized light.) In fact many organic substances in the body—e.g., the sugars, albumen, and lactic acid—are optically active. Such substances rotate the plane of polarization.[1] As a rule these substances have an absorption band a little to one side of the visible spectrum, and a curve of rotatory dispersion which increases with frequency. Some organic substances, such as chlorophyll and xanthophyll and others of the carotenoids, absorb frequencies within the limits of the visible spectrum. The result is that the curve of rotatory dispersion is also markedly anomalous. If, for example, the long wavelengths on one side of green are rotated toward the right, those on the short wavelength side will be rotated toward the left (33, pp. 494–495).

[1] See the *Glossary*, pp. 240–241, for definitions of this and other technical terms.

It is characteristic of the mixture of the complementary pairs that they cancel each others' hues and become achromatic. They lie on opposite sides of the curve of maximum sensitivity, which for the moment we will tentatively identify with the curve of absorption of the retinal substance. In addition to this orderly location with respect to the curve of maximum sensitivity, the pairs are "opposite." Some property of their respective frequencies can be plus or minus. Heretofore we have suggested nothing better as an explanation for this than that our imaginary primaries stimulate the respective cones equally.

We have now assumed that the complementary colors tell us something about the nature of the substance in the retina. This substance is achromatic for the various complementary pairs, because of the difference in the signs of their rotatory dispersion constants. There are no complements for hues lying in the middle of the spectrum, simply because as one approaches this region from the long wavelength side, the complementary frequency—i.e., one possessing a suitable rotatory dispersion constant—is beyond the visible spectrum on the short wavelength end. An approximation of this new curve is shown in Fig. 3.

The saturation of mixtures. The amount of a hue is expressed in terms of saturation, a term which Southall (28) believes ill-defined. And no wonder! According to Southall, "saturation is the degree in which the hue is undiluted by admixture with grey" (pp. 312–313). He expresses his dissatisfaction with this definition by saying: "If these definitions are necessarily vague, it is partly because hue, saturation, and luminosity, especially the two latter, are all more or less really interdependent to a certain extent" (p. 313). From the standpoint of laboratory procedures in controlling these variables, this appears to be the case. Again our operations are not definitive. In the laboratory the amount of hue can be altered by mixing white light with it, but the nature of saturation is not, thereby, illuminated.

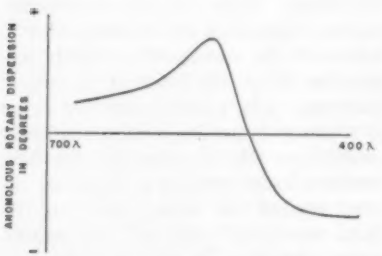


FIG. 3. An idealized curve of anomalous rotatory dispersion found in certain optically active substances.

There is, however, another way of desaturating a hue, and that is by mixing it with its complement. Now the question arises whether or not these two ways of desaturating a hue are the same. What are the relevant facts?

The frequencies in the spectrum can be separated with great purity. Yet it is commonly observed that the saturation of the spectral hues is greatest at the ends and least in the middle of the spectrum. The retinal process responsible for these differences in saturation cannot be a matter of white-light stimulation. But by the same reasoning, we cannot conclude that the "desaturation" resulting from the addition of the complement to a hue is *not* a retinal process. Neither can we argue that the same retinal process is subservient to both phenomena. There are indeed two separate events here, as we will attempt to show.

The curve of saturation has its peak slightly toward the long wavelength side of the maximum of absorption. A mixture of any number of hues seldom yields a hue as saturated as the most saturated member of the mixture,

whether these hues come from one or both sides of the spectrum. In no case does a mixture ever yield a hue which is greater in saturation than the most saturated hue of the mixture.

Moreover, the curve of saturation varies for different people. The range of variation is greatest for the color-weak and color-anomalous, according to Allen (1). For these people a shift in the point of minimum saturation down the spectrum toward the shorter wavelengths is accompanied by a loss in saturation over the whole of the spectrum.

Finally, all hues but red become completely desaturated near threshold. Try as we may, there appears to be no physical property of light which corresponds to saturation. Again we point to the fact that the physical properties of light accrue to it through its passage in a material medium. The first task, then, is to decide which of the above observations is the most important from the standpoint of leading us to a fruitful inquiry. As a beginning we will assume that the position of the peak of minimum saturation tells us something about the nature of the substance in the retina.

Our previous investigation led to the optically active compounds. In such substances plane-polarized light is split up into right and left circular components. Depending upon the frequency, if one of these components is absorbed a little more than the other (a condition known as circular dichrosm), the resultant beam will exhibit a certain amount of ellipticity depending upon the amount of the absorption. The point of maximum ellipticity occurs at a slightly lower frequency than the maximum of general absorption. The degree of ellipticity falls off toward either end of the spectrum. An approximation of the curve of ellipticities

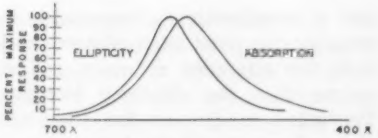


FIG. 4. Showing the relation of the curves of absorption and of ellipticity in certain optically active substances.

in relation to a curve of absorption is plotted in Fig. 4.

In its interaction with the retinal substance, the light has gained another property. This curve of ellipticity bears closer scrutiny than the mere appreciation of its position with respect to the position of the curve of saturation. With a piece of graph paper the interested reader can easily show that the combining of two ellipses, even when assumed to have opposite rotations, will give a third ellipse. In other words, the curve of ellipticity could very well subserve the saturation function.

There is no way of controlling the amount of ellipticity of a given frequency. Neither can we alter any property of a given frequency in such a way as to affect its saturation. In a mixture one can get less, but not more; the result is an accident of the frequencies mixed. For a mixture of the complementary hues, one can simulate the control of saturation through the adjustment of the intensity of one or the other members of the pair. With noncomplements this will not work, since a change of intensity in one or the other of the components will alter the hue of the mixture without materially affecting the saturation. Hence, in a mixture of two or more frequencies no two of which are complementary, the resultant mixture will not be quite so saturated as the hue mediated by the single frequency which it matches.

In summarizing the above discussion it would appear that as the one member from the long wavelength side of each

pair of complementary frequencies approaches the point of maximum sensitivity the saturation decreases, and we assume that the ellipticity increases. The allied member on the other side of maximum sensitivity becomes more saturated, and again we assume that the ellipticity which is of opposite sign decreases. The cancelling takes place through opposite dispersions, and for this reason the hue will be one or the other, depending, first, on the respective intensities, and secondly, upon the amount of ellipticity. Since the ellipticity is fixed for a given wavelength, the intensity must be adjusted in order to offset both the effect of the differences in dispersion and the amount of ellipticity, to obtain complete achromaticity.

On the presence of unsaturated compounds in the eye. There is no lack of optically active substances in the eye (30, 32). B-carotene, vitamin A, and xanthophyll are examples. The first of these is an optically active compound having the general formula $C_{40}H_{56}$. It possesses eleven double bonds in a *trans* arrangement, five of which allow rearrangement to form some twenty stereoisomeric forms. While these stereoisomers have the same formula, they not only differ in their physical constants, but also possess different absorption spectra. According to Wald (30), B-carotene and vitamin A show sharp narrow absorption bands at 618 and somewhat broader ones between 400 and 500 mu. The former shows a specific rotatory power [a] 364 degrees for the line Cd. All of the stereoisomers of B-carotene have absorption bands farther down the spectrum toward the shorter wavelengths.

B-carotene is significant in that it shows the presence of an optically active substance in the eye. It is quite plausible, however, for one of the various other carotinoids to be present. These range from phytofluene, $C_{40}H_{64}$, which has no absorption band in the visible spectrum, to torulene, the absorption band of which is shifted toward the red end of the spectrum. Torulene has 13 double bonds, and can form even more stereoisomers than B-carotene, one of which could have an absorption band in the yellow-green. There are still other possibilities. Iodosin has been identified and proposed as the cone substance responsible for color vision (2). The substance sought for, however, may come into existence only upon exposure of the retina to light.

An account of hue, brightness, and the photochromatic interval. In the absence of any certain knowledge of what the cones contain (6), we might postulate that they contain a carotinoid described by the general formula $C_{40}H_{56}$. Under the action of light, decomposition is not indicated, since (*a*) no heat is produced; (*b*) light adaptation is very rapid; and (*c*) the greater part of dark adaptation is also very rapid.

The following is tentatively proposed as a means of redirecting our present thinking along new lines. It is clearly recognized by the author that the account may be oversimplified and possibly incorrect.

Under the influence of light $CO_2 + H_2O$ combine to give $CH_2O + O_2$. Now it would appear that O_2 is being liberated. It is more likely, however, that as the CH_2O forms it acts as a catalyst for the $CH_{40}H_{56}$, which, in the presence of light (*a*) acts as a hydrogen donor for the excess O_2 to reform H_2O, and (*b*) has the number of its double bonds increased by giving up four H atoms to the O_2, resulting in the formation of $C_{40}H_{52}$. Such a sequence of events and the resultant stereoisomers would be of the utmost importance, because this dehydrogenation (36) would be followed by a broadening of the absorption spec-

trum and a shift in the point of maximum absorption toward the longer wavelengths. The process reaches a state of equilibrium when the excess CO_2 is used up. At this point we might say the eye is prepared to see.

Upon exposure to any of the various light frequencies, a unique set of events occurs. The new substance has a certain number of free double bonds, i.e., ones around which different arrangements of the atoms can occur. When a certain rearrangement takes place, it renders the molecule temporarily inactive to light of that particular frequency. The new arrangement exists for only a short time before it snaps back into its original form. In doing so energy is given up. It is not only the absorption of energy which excites the cone, but also the release or change in energy. Hence neural impulses are set up both when light falls on the cone and also when it is removed.

All cones are alike in that they contain the same substance. They differ, no doubt, in the amounts they contain; this fact, together with the additional one that the refractile tips also vary somewhat in length, probably renders the individual cone differentially sensitive to narrow bands of light. This may be in part what Granit discovered. On the whole, however, the cones act alike. As the new substance comes into being during the initial exposure, the cone is excited. With greater levels of illumination greater numbers of molecules are prepared to function in differentiating between the frequencies imposed upon them. At threshold levels of illumination this chemical process underlies the photochromatic interval. It is also responsible for brightness adaptation. A careful distinction should be made between this and the production of stereoisomers, which is the basis for color vision and adaptation to hue.

Much progress has been made toward understanding how the rods work. We are all familiar with Hecht's account of the conversion of rhodopsin under the action of light into retinene and vitamin A. On this subject Wald has written extensively. Recently he has extended his account for the purpose of including the cones also (32).

Wald contends that each rod has many compartments containing rhodopsin. Only one quantum is necessary to activate a compartment. This quantum is absorbed by one molecule. Although the other molecules in the compartment each absorb a quantum, they contribute nothing to the response of the compartment. On the contrary, they slow down the process of dark adaptation because the compartment cannot become active again until the last molecule of rhodopsin has regenerated.

The fact that each molecule of rhodopsin absorbs only one quantum of light, Wald contends, is very important for understanding how the cones are activated. The cone contains only a small fraction of the amount of rhodopsin contained in the rod. It has, therefore, a high threshold. He suggests that if we assume that if each compartment of the cone contains only one molecule, the rapid rate of recovery during dark adaptation would be accounted for.

Wald's analysis raises certain questions in the mind of the writer, the answers to which are not immediately apparent. We must assume that at threshold the neural effect is equal in amount for both rods and cones, at all frequencies. From the work of Chapanis (4), however, it seems that with respect to the cones more quanta are necessary for the yellow threshold than for the red or blue. As to the rods, more quanta of red are necessary than for yellow or blue. How can this be if the substance is the same?

In the present account we prefer to

think that the active substance in the cones is not the same as that in the rods. This does not exclude the possibility that the cones contain some rhodopsin. The account is interesting in that it provides one means of accounting for brightness in color perception. It does not account for the perception of hue, nor, indeed, was it intended to.

In the present account, the molecule of our hypothetical substance is analogous to Wald's compartment. The fact that the structure of the molecule can be rearranged also accounts for hue.

The color-anomalous. This great field of study deserves more detailed attention than the present brief account can spare. Our general ideas about the color-weak and the color-blind have given rise to certain misconceptions, both as to their being classified into distinct classes and as to what these people can see. There are no separate classes (23). Rather the color-anomalous deviate from the normal by differing amounts, the smaller deviations being more numerous. But because of their relative unimportance in everyday practical affairs, they become lost to our attention.

Again, such terms as red-green blindness and yellow-blue blindness obscure the fact that inability to see red may result from a set of factors different from those involved in inability to see green. Some red-blind show low sensitivity to the long wavelengths. They are insensitive to red because their maximum luminosity is shifted toward the green. They do not see red as yellow. The normal orangish reds and oranges are visible as very desaturated yellow. In fact, the whole of the spectrum is usually desaturated. Blue-green is not seen, because it is greatly or completely desaturated near 496. When both red and blue-green are of equal brightness they are confused, but they are equally mistaken for a gray. Other red-blind have normal sensitivity for the long wavelengths. They are green-blind because the greens are completely desaturated near 504.

The yellow-blue-blind are still more complicated. This type is sensitive to red, but the yellow is completely desaturated. The luminosity curve for the tritanope is fairly normal. The point of minimum saturation is about 565 or 570, with another in the extreme violet. Still others have points of minimum saturation at 580 and 470 (1). In the Richard-Robinson case (27), the subject was extraordinarily weak to green.

Optically active compounds often show two regions of absorption. When these fall within a certain range, the curves of anomalous rotatory dispersion and of ellipticity are considerably altered. It is not necessary for the second absorption point to fall within the visible spectrum. The explanation for the color-anomalous suggested here assumes that small chemical variations are present, probably in the form of minute amounts of slightly different substances. Note, for example, that B-carotene has the general formula $C_{40}H_{56}$. It is found in a slightly different form in xanthophyll $C_{40}H_{56}(OH)$, $C_{40}H_{56}(OH)_2$, and $C_{40}H_{56}(OH)_3$.

This possibility is especially significant because the curve of anomalous rotatory dispersion can be produced in different ways. The one tentatively selected here involves only one kind of molecule. Anomalous rotatory dispersion can, however, be produced by different molecules in the same solution, and again by different molecules in different solutions. It seems unreasonable to the writer, in the light of what we know about the complex chemistry of the retina, not to expect such variations. In fact, our anticipation of them not only provides an explanation of the color-blind, but strengthens considerably our contention that the normal

trichromat, the red-green blind, and the yellow-blue blind are not separate classes.

The afterimage. A sensible account of the afterimage continues to pose one of the most difficult problems in the long history of color vision. Judd's brief account in Stevens' *Handbook* (16) clearly shows that this difficulty is primarily one of trying to resolve the many apparent inconsistencies by portioning them out among the various inadequate theories.

Helson's (14) work shows that, upon adapting the whole retina, exposure to a surface of greater reflectance results in perception of the adapting light. Upon exposing the adapted eye to a surface of lesser reflectance, the complement is perceived. These findings are chiefly of interest in demonstrating that the afterimage is not produced by incoming light, following adaptation. Further support of this is found in Southall's report (28, p. 343) that, upon adapting the eye to a deep red, followed by exposure to a yellow, the latter looks greenish-yellow; whereas this same yellow looks orange when the adapting light is blue-green. Lastly the complementary afterimage develops in complete darkness.

These observations appear to place difficulties in the way of accepting Woodworth's modification (34, pp. 558–561) of Ebbecke's cortical theory. Moreover, the greatest saturation is experienced when the afterimage is projected upon a background of the same hue. In Woodworth's theory, the cortical effects should be opposed. Theoretically, then, the hue should be less bright and presumably less saturated, whereas in fact neither of these occurs.

That a hue becomes less bright and less saturated after gazing at it for some time is common experience. And who has not seen in the cloud of his favorite pipe the genie of the afterimage? That the physiological events subserving these phenomena occur simultaneously with exposure cannot be denied. How remarkable indeed would it be that chance alone should fix the hue of the afterimage, so developed, so nearly to the complement.

Again, is it an oversight that our present theories offer no explanation for the increased saturation that appears when the complementary afterimage is viewed against a background of the same hue? Expressed in a different way, note that adaptation to red makes the eye "red-blind," but at the same time simultaneously increases the blue-green sensitivity. Immediately we are challenged not to permit our explorations to be channelized by traditional modes of thinking. Helmholtz and Hecht ignore the afterimage; Granit and Hartridge are unconcerned. Hering stressed its importance, as did Ladd-Franklin, but the force of their arguments has only succeeded in bringing about an impasse.

We do not believe the above events are unrelated accidents. Combined, they tell us something about the action of light upon the retinal substance. The molecule of such a photochemical substance, for example, B-carotene, has a certain number of double bonds around which stereoisomers are formed. These particular double bonds we may think of as elastic couplings, because the rearrangements brought about by the action of light revert to their original form within a short time.

Here we must digress for a moment to study the meaning of the term "absorption." As the psychologist usually uses it, it refers to the transformation of light energy into heat energy. The light which acts upon the retina is not absorbed in this sense. Depending upon the substance, Wood points out the many ways light energy is disposed of. For example, it may be transmitted, reflected, absorbed (i.e., changed into

heat), reradiated (concurrently or later), or, as in photosynthesis, converted to a different kind of energy and stored. In the case of a substance like xanthophyll, energy is momentarily converted to potential energy and held in the new arrangements, although certainly some portion is lost in various other ways. About ten to twenty per cent of this substance bleaches during a period of an hour.

Special mention should be made of that form of reradiation known as fluorescence. This phenomenon, which occurs when the blood and nerve fibers are exposed to certain frequencies, we want carefully to avoid, since it contributes nothing to our account. In the following discussion, then, fluorescence is not considered.

The physiological account of adaptation, which is basic to the understanding of the afterimage, is qualitatively fairly simple. It depends for the most part upon the amount of available substance, the rate of transformation, and the rate of recovery to the original state. It should be clear that, since these transformations are physical events, not chemical ones, there is no question of fatigue involved here and no necessity to consider nutrient replacements. But because of this fact it is necessary to ask what becomes of the energy which the rearrangement gives up when returning to its original form.

Two suggestions immediately present themselves. The first concerns direct reradiation, in which the energy released would find its way into the surrounding medium to be reabsorbed or dissipated in various other ways. This cannot be what happens, since surely the illuminated area would gradually grow larger and larger, instead of showing, as it does, a small edge of complementary hue surrounding the test spot. Moreover, there are certain losses due to frictional damping, resulting in smaller quanta being reradiated. Hence the reradiated energy would necessarily consist of a lower frequency than the original light, and the areas away from the test spot on the retina would be filled with bands of different hues, corresponding to longer and longer wavelengths.

The very form of the arrangements of the atoms in the molecule which accounts for the curve of anomalous rotatory dispersion suggests another possibility. The forces created by the return of a certain rearrangement to its original form generate equal but opposing forces in some other part of the same molecule, resulting in a shift and a storing of energy in another rearrangement. As a physical system, the loci of such shifts are always the same and completely reciprocal.

This brief account concludes our development of what we believe is a consistent account of color vision. We cannot, in this proposal, settle certain metaphysical questions relating to the concept of experience and to the larger mind-body problem from the traditional point of view. In review, we can merely restate our position as one which envisages the mechanism of visual contact as a group of interactions between light energy and the substances in the eye. When such interactions are understood and properly developed, we believe that it will be unnecessary to burden the nerve fibers with any other function than that of conducting impulses, regardless of their position in the body.

Afterthoughts and further problems in vision. No one scientist, working alone, can develop a sensible theory of vision. It must first be determined whether the present suggestions regarding the properties of "visible" frequencies, as these properties accrue to the light rays in their passage through an optically active compound, are reasonable. If they are, the resultant change in the optically active substance permits a thoroughly

objective interpretation of vision. The implications, if such a theory were accepted, could be far-reaching, since much of our psychology has been greatly influenced by visual theory. Impeding our progress in this direction are many facts which need reinterpretation.

Examine, if you will, the following simple example that fosters certain of our present misconceptions. In physics a whole chapter may be devoted to the formation by lenses and mirrors of "images." No ingenuity whatever is needed to demonstrate that a surface of some kind is necessary for the "image" to fall on in order for it to be seen. At this point the analogy between the eye and the camera is usually pointed out. But this analogy is peculiarly deceptive. The camera is a passive inert thing. It has no behavior of a manipulatory character. The chemicals of the film are altered but the camera does no seeing.

The lens placed between the eye and an object becomes part of the media of contact between the two. A paper placed at the focal point of the lens also becomes a part of such media. It allows the observer to look around corners. The paper alone, because of diffusion, distorts the media. A plane mirror would not. An imperfect lens does not distort the image; it distorts the total media. But in any case one sees not an "image." One is simply placed in contact with the object.

The question is not one of semantics nor one of "point of view." The record on a film is not an image; it is an arrangement of chemical particles, an ingenious method set up by the experimenter by which the object reproduces its likeness. The labor of arrangement (manipulation) is done beforehand. The experimenter has at his hand a light which he fixes just so. This light is exactly analogous to thousands of levers all poised and ready. The moment the experimenter turns on his light these thousands of levers create a copy by prying and pushing in one fell swoop, coupling each point of the object from which they are guided by the lens (fulcrum) with the particles on the film. The labor done beforehand is the fixing of the silver particles in such manner that the light levers which put the film into contact with the object also record this fact.

Whether the organism looks directly at an object, or through a lens, or via a piece of white paper at the focus of the lens, is all the same. There are many other gadgets—reduction tubes and screens are examples—which will affect contact by their action upon the light. They all become part of the media of contact.

No thought is entertained here concerning the association of these properties or their effects upon the molecular structure, with impulses either in particular sense cells, nerve fibers, or areas of the brain. No temporal or spatial patterning of impulses is called upon to supply any of the so-called secondary qualities. Again, no "special ability to discriminate between different receptor impulses" needs to be "developed in the central nervous system" (6, p. 41). The point of view offered here emphasizes the limitations placed upon theory when specialized cells are considered as "receptors." Particular energies do not affect peculiar "qualities" either in the receptor, in the attached nerve fiber, or in the area of the brain to which this fiber leads. The problems with which Müller struggled, and to which his solutions were so readily accepted by Helmholtz, are here considered as most naive in the light of present knowledge. We do not feel that Ladd-Franklin was at all helpful when, in spite of her logic, she suggested that ". . . if one fastens his mind upon the color-triangle (the map of possible color-sensations), and forgets the visible spectrum, compre-

hension of the phenomena of color will be greatly facilitated" (20, p. 200).

Constructively, the present theory points to the light frequencies and the associated phylogenetically developed substances in the eye as media of contact between objects, possessed of certain properties by reason of their interaction with light, and the eye as a biologically irritable structure.

Concern over differential nervous effects is neither warranted nor welcomed. Within certain limits such effects do occur. Contact with the various separable properties of objects is correlated with differential effects in the organism, but it is argued that these differential effects are alterations in the organism's chemistry. It is further argued that these interactions must obtain before an organism can perform an act of discrimination considered as an elementary manipulable.

The organism perceives visually only after contact with objects is followed by integrated manipulation of the object, in and of itself or in relation to its setting. Thus perception has come by reason of its complexity and usefulness to win our present efforts, and psychophysics, the study of sensing, has fallen into disrepute—this largely because its claims were too great. Molecules may be built of atoms, houses of brick, pounds of dynes, and weeks of days; but not behavior of reflexes or visual perceptions of rods, cones, fibers, and brains.

The house of brick only vaguely portrays my labor, my ambition, my compromises, my values in gain and loss. Sensing is still a useful concept to denote that contact with a real world is mediated by energy in different forms, according as the biological organism has adapted phylogenetically to the world of things and events within which it has developed.

GLOSSARY

Carotene: Any of a class of optically active yellow, orange, and red substances having the general formula $C_{40}H_{56}$.

Dichrosm, circular: The unequal absorption of right and left circularly polarized light by a medium in which the optical activity and the absorption are produced by the same molecule.

Dispersion: The divergence of the pathways of the different frequencies as they travel through a refracting medium.

Dispersion, anomalous: A rearrangement of the order of hues. In absorbing media the velocity of a frequency on the long wavelength side travels more slowly as one approaches the maximum of the absorption band. Approaching the region of maximum absorption from the short wavelength side, the frequencies travel faster. This results in a rearrangement of the order of the hues in a spectrum produced by a prism of such a substance.

Dispersion, rotatory: The variation in the degree of rotation of the plane of polarization with frequency.

Dispersion, anomalous rotatory: The variation in the degree of rotation of the plane of polarization with frequency, which occurs when an optically active absorption band falls within the range of visible frequencies.

Optical activity: Rotation of the plane of polarization.

Polarization, plane: An arrangement of vibrations from a ray of light. It is plane-polarized when all of the vibrations occur along a given line in a plane at right angles to the direction of propagation.

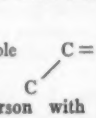
Polarization, circular: The circular vibration of a ray of light brought about by passing plane polarized light through a quarter-wave plate, a Fresnel romb and many organic substances.

Polarization, elliptical: The elliptical vibration of a ray of light. It is exhibited by substances which show circular dichrosm, but can be produced in many other ways.

Polarization, rotatory: Rotated plane of polarization. When a plane-polarized ray of light is passed through a circularly double refracting medium, the resulting plane of polarization will be rotated to the right or left when compared to the incident ray.

Refraction, index of: The ratio of the velocity of light in a given substance to the velocity of light in air.

Stereoisomers: Compounds which differ only in the spatial arrangement of their atoms. Their chemical formulae are the same.

Transconfiguration or arrangement: A stereoisomer whose form resembles a Z on its side, for example  .

Trichomat: A person with normal color vision, i.e., one who needs three selected hues to mix together in creating a match for any of the colors.

Tritanope: A person who is partially color-blind; specifically, one who is relatively insensitive to blue or who confuses blue and yellow.

REFERENCES

1. ALLEN, F., ARCHIBALD, D. C., & LIND, R. A. The coordination of normal and abnormal color vision. *J. opt. Soc. Amer.*, 1929, **18**, 1–16.
2. BLISS, A. F. The chemistry of daylight vision. *J. gen. Physiol.*, 1946, **29**, 277–297.
3. CHAPANIS, A. Spectral saturation and its relation to color-vision defects. *J. exp. Psychol.*, 1944, **34**, 24–44.
4. CHAPANIS, A. The dark adaptation of the color anomalous measured with lights of different hues. *J. gen. Physiol.*, 1947, **30**, 423–437.
5. FORBES, W. T. M. An interference theory of color vision. *Amer. J. Psychol.*, 1928, **40**, 1–25.
6. GOTHLIN, G. F. The fundamental colour sensations in man's colour sense. *Kungl. Svenska Vetenskapaskademiens Handlingar*, 1943, **20**, No. 7. Pp. 76.
7. GRANIT, R. The color receptors of the mammalian retina. *J. Neurophysiol.*, 1945, **8**, 195–210.
8. GUILD, J. The colorimetric properties of the spectrum. *Philos. Trans.*, 1931, **A 230**, 149.
9. HARDY, A. C. *Handbook of colorimetry*. Cambridge, Mass.: Technology, 1936.
10. HARTRIDGE, H. Recent advances in color vision. *Science*, 1948, **108**, 395–404.
11. HECHT, S. On the binocular fusion of colors and its relation to theories of color vision. *Proc. Nat. Acad. Sci.*, 1928, **14**, 237–241.
12. HECHT, S., SHLAER, S., & PIRENNE, M. H. Energy at the threshold of vision. *Science*, 1941, **93**, 585–589.
13. HECHT, S., SHLAER, S., & PIRENNE, M. H. Energy at the threshold of vision. *J. gen. Physiol.*, 1942, **25**, 819–840.
14. HELSON, H., & JEFFERS, V. B. Fundamental problems in color vision: II. Hue, lightness and saturation of selective samples in chromatic illumination. *J. exp. Psychol.*, 1940, **26**, 1–27.
15. JUDD, D. B. The 1931 I C I standard observer and coordinate system for colorimetry. *J. opt. Soc. Amer.*, 1933, **23**, 359.
16. JUDD, D. B. Basic correlates of the visual stimulus. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951. Pp. 811–867.
17. LADD-FRANKLIN, CHRISTINE. Color introspection on the part of the Eskimo. *Psychol. Rev.*, 1901, **8**, 396–402.
18. LADD-FRANKLIN, CHRISTINE. Special review of J. H. Parsons, *An introduction to the study of colour vision*. *Psychol. Bull.*, 1916, **13**, 272–275.
19. LADD-FRANKLIN, CHRISTINE. On color theories and chromatic sensations. *Psychol. Rev.*, 1916, **23**, 237–249.
20. LADD-FRANKLIN, CHRISTINE. Practical logic and color theories. *Psychol. Rev.*, 1922, **29**, 180–200.
21. LADD-FRANKLIN, CHRISTINE. The nature of the colour sensations. In *Helmholtz's treatise on physiological optics* (English translation, James P. C. Southall, Ed.). Optical Society of America, 1924.
22. LOWRY, T. M. *Optical rotatory power*. London: Longmans, Green, 1935. Pp. xii, 483.
23. MILES, W. R. Color blindness in eleven thousand museum visitors. *Yale J. Biol. & Med.*, 1943, **16**, 59–76.
24. PRIEST, I. G. Preliminary note on the relations between the quality of color and the spectral distribution of light in the stimulus. *J. opt. Soc. Amer.*, 1920, **4**, 388–401.
25. PURDY, D. McL. On the saturations and chromatic thresholds of spectral colours. *Brit. J. Psychol.*, 1930–31, **21**, 283–313.
26. RICH, G. J. An eclectic theory of vision. *Psychol. Rev.*, 1928, **35**, 311–318.
27. RICHARDSON-ROBINSON, F. A case of color blindness to yellow and to blue. *Amer. J. Psychol.*, 1923, **34**, 157–184.
28. SOUTHALL, J. P. C. *Introduction to physiological optics*. New York: Oxford Univer. Press, 1937. Pp. x, 426.
29. VAN NIEL, C. B. Comparative biochemistry of photosynthesis. *Amer. Scientist*, 1949, **37**, 371–383.
30. WALD, G. Carotenoids and the visual cycle. *J. gen. Physiol.*, 1935, **19**, 351–371.

31. WALD, G. On the mechanism of the visual threshold and visual adaptation. *Science,* 1954, **119**, 887–892.
32. WALD, G. The molecular basis of visual excitation. *Amer. Scientist,* 1954, **42**, 73–95.
33. WOOD, R. W. *Physical optics.* London: MacMillan, 1919. Pp. xvi, 705.
34. WOODWORTH, R. S. *Experimental psychology.* New York: Henry Holt, 1938. Pp. xi, 889.
35. WRIGHT, W. D. *A re-determination of the trichromatic mixture data.* London: Medical Research Council, Report of the Committee on the Physiology of Vision, VII, Special Report Series, 1929, No. 139.
36. ZECHMEISTER, L. Chromatography and spectroscopy in organic chemistry and stereochemistry. *Amer. Scientist,* 1948, **36**, 505–516.

(Received August 8, 1955)

Psychological Review
Vol. 63, No. 4, 1956

A FIELD THEORY OF EXTINCTION AND SPONTANEOUS RECOVERY

KOH JOHDAI

Hiroshima University

When a rat in a Skinner box (A) tries to make a lever-pressing response (L) in order to reach food (G), we can make the following statement about the direction of a psychological force $f_{A,G}$ in the field: $d_{A,G} = d_{A,L}$. But when the rat no longer tries to press a lever after a certain number of nonrewarded responses, we can make another statement about the direction of the force at that moment: $d_{A,G} \neq d_{A,L}$. In other words, after withholding of food, extinction of lever-pressing response of a rat in a Skinner box can be thought of as due to the change of a force $f_{A,G}$ in direction from $d_{A,G} = d_{A,L}$ to $d_{A,G} \neq d_{A,L}$.

The possibility that extinction may be due to the changed direction of a force $f_{A,G}$ has not been examined, however, because extinction procedures used in the experiments up to this time have dealt merely with the conditioned response to be extinguished: there exists no region Y such that $d_{A,G} = d_{A,Y}$. If we provide a situation where an alternative response (Y) signifies extinction of the original conditioned response (X), we may be able to relate extinction to the changed direction of a psychological force: change of the force $f_{A,G}$ in direction from $d_{A,X}$ to $d_{A,Y}$ can be brought to light.

The procedure for extinction through the somewhat new "response-alternation" method employed in a preceding study (**4**) was designed to test the above thesis. Human subjects were required to move a stylus without sight from the starting point A to the same point along the path $w_{A,E,G}$ in a maze shown in Fig. 1 (redrawn from the original report) during the acquisition period of 16 trials. Then, in the extinction period, resistance to extinction was measured by the number of nonreinforced locomotions along the path $w_{A,E,G}$, which were replaced by

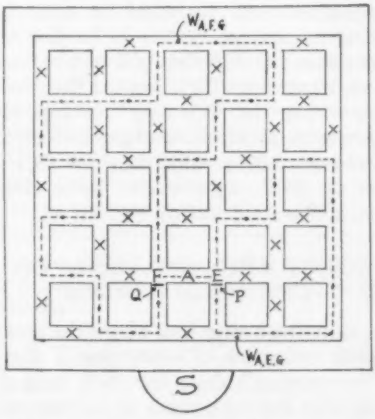


FIG. 1. Pattern of the stylus maze used. A, starting point at one time and terminal point at another time. The starting was indicated by an ordinary electric light, and arriving at the goal (G) was indicated by a red light which meant scoring 20 points (reinforcement). The experimenter could switch on at will either of these two lights whenever the stylus was placed on a point A. He also could prevent them from being lighted even when the stylus was placed on a point A (nonreinforcement). Whenever the stylus was brought to a point E, a green light was exposed at the window; whenever the stylus was brought to a point F, an orange light was exposed. P and Q indicate one-way doors which permitted locomotions only in the direction of arrows. Whenever the stylus was brought to a point X, S was given a buzzer shock. S indicates location of S; $w_{A,E,G}$, green path; $w_{A,F,G}$, orange path.

the locomotion along the path $w_{A,F,G}$ at the moment of complete extinction. That is, this procedure is characterized by the fact that complete extinction of a partially or continuously conditioned response is the making of another response for the same goal.

According to Miller and Dollard, "the apparent function of extinction is to eliminate responses which do not lead to reward, so that other responses can occur" (**7**, p. 38), and "the function of extinction is to force the subject to try new responses" (**7**, p. 43). Extinction in such an experimental situation may therefore be said, in field-theoretical terms, to be due to the changed direction of a two-valued resultant force $f^*_{A,G}$ toward the goal, assuming the following hodological-geometrical definition of psychological directions (**6**, p. 46): $d_{A,E,G} \neq d_{A,F,G}$ if no path $w_{A,X}$ exists such that $w_{A,X} \subset w_{A,E,G}$ and $w_{A,X} \subset w_{A,F,G}$.

RELATIVE EFFORTFULNESS OF THE CONDITIONED RESPONSE

Suppose a situation where the complete extinction of a response X (i.e., "locomotion" along the path $w_{A,E,G}$) signifies the occurrence of another response Y (i.e., "locomotion" along the path $w_{A,F,G}$). In this case, we may call into question the *relative effortfulness* between the two responses X and Y. Since a number of experiments on extinction carried out up to this time called into question only a response to be extinguished, it was natural that experimenters dealt with merely the amount of effort required for *the* response (k)[1] and not with the *relative effortfulness* between the two responses (rk) in our sense. Rather they dealt with the case of extinction where the alternate response (Y) was impossible (in other words, it had infinite effortfulness). For example, a rat in the Skinner box has no other means at the time of complete extinction of its lever-pressing response. Thus, in the conventional extinction procedure, only the *boundary case* of extinction was treated. In such a case of extinction, for example, the bar load (**8**) and the distance to be jumped (**10**) as the indicators of effortfulness of the conditioned response are the main determiners of extinction. Therefore, the resistance to extinction (n^e) may be represented as a function of the effortfulness of the response to be extinguished (X):

$$n^e = F(k_{A,E,G}), \qquad [9][2]$$

where $k_{A,E,G}$ means the total effortfulness of a locomotion along the path $w_{A,E,G}$. In our case, on the other hand, the resistance to extinction of a response X may be viewed as a function of the *relative effortfulness* of X against Y: the resistance will vary with the increasing or decreasing effortfulness of Y, even when the effortfulness of X is kept constant. Thus it may be that

$$n^e = F(rk_{A,E,G}). \qquad [9a]$$

Extinction in this case is viewed as due to the change of the force $f_{A,G}$ in direction from $d_{A,E}$ to $d_{A,F}$.

The response Y, whose occurrence signifies the complete extinction of the response X, tends to be replaced by X, in turn, even when successively reinforced, especially in cases where the response Y is much more effortful for the subject than X. This phenomenon is nothing less than a spontaneous

[1] Lewin (**6**, p. 153) formulated by the equation $k_{A,G} = F(e_{A,G}, av_{A,G})$ that the effortfulness of a response (locomotion) is a function of the distance (e) and the average velocity (av). For the present, we shall treat "k" in the sense of the formula as well as for difficulties of the task.

[2] Equations in this paper are numbered in sequence with those in the preceding article (**4**).

recovery (*Re*). In such a case, the successively reinforced trials of the response *Y* may be viewed as "interpolated activity" which will affect the spontaneous recovery of *X* in this way: the greater the effortfulness of response *Y*, the stronger the tendency to spontaneous recovery of *X* after a smaller number of trials of the response *Y*. Thus it may be that

$$Re = F(rk_{A,E,G}). \qquad [10]$$

Spontaneous recovery in this case is viewed as due to the change of the force $f_{A,G}$ in direction from $d_{A,F}$ to $d_{A,E}$.

Hypothesis 10 has been partially verified by the experiment with human subjects, although, as far as the difference in relative effortfulness of *X* given in the experiment is concerned, it has not been accompanied by the results for which Equation 9a is adequate (**5**). These results, however, seem to mean that *spontaneous recovery is more sensitive than extinction in respect of the relative effortfulness of the responses.* No significant difference has been found between the effects of different relative amounts of effort upon extinction, in spite of the fact that they have been accompanied by statistical significant difference in the tendency to spontaneous recovery. Greater differences in the relative effortfulness of the response *X* would prove our Equation 9a as well as 10.

Thus, *both extinction and spontaneous recovery can be viewed hodologically as due to the changed direction of the final resultant force* $f^*_{A,G}$ *in the field.*

DECISION IN AN OVERLAPPING CHOICE SITUATION

On considering extinction and spontaneous recovery as such, then, we must deal with two resultant forces:

$$f_{A,E,G} + f_{k_{A,E,G}\cdot -k_{A,E,G}} = f^*_{A,E,G}$$

and

$$f_{A,F,G} + f_{k_{A,F,G}\cdot -k_{A,F,G}} = f^*_{A,F,G}. \qquad [11]$$

The two components of each resultant force are, of course, opposite in their direction to each other. Thus, just as in a situation of the obstruction box (**6**, Ex. 13, p. 76) or in measuring forces by velocity of locomotion (**6**, pp. 154–155),

$$d_{A,E,G} = \bar{d}_{k_{A,E,G}\cdot -k_{A,E,G}}$$

and

$$d_{A,F,G} = \bar{d}_{k_{A,F,G}\cdot -k_{A,F,G}}, \qquad [11a]$$

where $\bar{d}$ indicates the opposite direction in a hodological space.

The effortfulness of the response *X* is usually, as in the preceding experiment (**4**), smaller than that of the alternate response *Y* ($k_{A,E,G} < k_{A,F,G}$). Thus, just as in the choice between two (longer and shorter) paths to the same goal (see **6**, Fig. 48, p. 169), the fact that every subject took the path $w_{A,E,G}$ during the training period in the preceding study (**4**) seems to be due to

$$f^*_{A,E,G} > f^*_{A,F,G}, \qquad [12]$$

even if $f_{A,E,G} = f_{A,F,G}$ (see **6**, Ex. 37, p. 168).

Further, the response alternation at the time of complete extinction means a "decision" in an overlapping choice situation (**6**, p. 206); in other words, since

$$|f^*_{A,E,G}| \cdot Po(w_{A,E,G}) = 0$$

and

$$|f^*_{A,F,G}| \cdot Po(w_{A,F,G}) = |f^*_{A,F,G}|$$

because

$$Po(w_{A,E,G}) = 0$$

and

$$Po(w_{A,F,G}) = 1, \qquad [13]$$

at least at some moment of such a decision, the occurrence of a locomo-

tion along the path $w_{A,F,G}$ (velocity of locomotion $v_{A,F,G} > 0$) may be viewed as due to the resultant force $f^*_{A,F,G}$ (see **4**, Equations 6 and 7).

It follows that spontaneous recovery in our type of experiment is due to the changed direction of a psychological force from $d_{A,F,G}$ to $d_{A,E,G}$. Spontaneous recovery (*Re*) implies that

$$|f^*_{A,E,G}| \cdot Po(w_{A,E,G}) > |f^*_{A,F,G}| \cdot Po(w_{A,F,G}) \quad [12a]$$

under the condition of $d_{A,E,G} \neq d_{A,F,G}$.

MEASURES IN ACQUISITION AND EXTINCTION

Errors, time, and latency have been the usual measures for both acquisition and extinction. Even though these measures in both processes are quite similar in *operational or empirical* characteristics, from our viewpoint these measures do not have the same psychological meaning (i.e., logical and conceptual properties in terms of scientific constructs) for acquisition as they have for the process of extinction. Increments in the process of extinction do not seem to indicate a simple retrogression of learning (delearning) to the earlier stage of acquisition. *Phenotypical or directly observable similarities do not always guarantee genotypical or dynamic ones* (**4**, footnote 7).

The increased latency in extinction seems to be due to the *undecided direction* of a psychological force. On the other hand, long latencies, especially at the beginning stage of acquisition, seem to be due to the *cognitive unstructuredness* related to the starting situation and the *low degree of unity of the beginning step of the path.* As the cognitive structurization and increment in the unity of the step are usually gradual when animals (e.g., rats) are used as subjects, the latency curve obtained represents the acquisition process as gradual. When human beings, especially adults, are used as subjects, however, the results are different (**5**, Fig. 4). This seems to be due to the fact that in human subjects cognitive structurization and unity of the beginning step, the main factors which determine the latency curve in acquisition, become asymptotic after only one or two trials.

In the process of extinction, the degree of cognitive structuredness and unity of the beginning step seem to remain high even right up to the complete extinction (see **4**, U_A and U_B in Fig. 3). Latency in extinction in an overlapping choice situation, therefore, must be another factor which might be termed "decision time."

The usual increment in the time required for response in the extinction process (**5**, Fig. 3) seems to be due to the two following factors: (*a*) the *decreased relative potency* (*Po*) *of the path* $w_{A,E,G}$ during the period of locomotions along it. As Equation 13 explains, the decreased potency of the path $w_{A,E,G}$ is necessarily accompanied by the weakened force along the path. The weakened force, in turn, is necessarily accompanied by the decreased velocity in spite of a high degree of unity of the path, as indicated by the equation

$$v_{A,E,G} = F(|f^*_{A,E,G}|, uni(w_{A,E,G})); \quad [14]$$

(*b*) the *increment of the strength of the negative valence of the region* $k_{A,E,G}$ ($Va(k_{A,E,G}) < 0$). Thus, at the time of complete extinction,

$$|f_{A,E,G}| < |f_{k_{A,E,G}, -k_{A,E,G}}|. \quad [11b]$$

This causes the subject to locomote along the path in the direction $d_{A,-k_{A,E,G}}$, that is, in the direction $d_{A,k_{A,F,G}}$. The factor *a* corresponds to the one which determined the "giving up" of the older children in

Fajan's experiment. "The infants, on the other hand, withdraw mainly as a result of an increasing negative valence of the barrier" (6, p. 130), i.e., due to Factor *b*.

Increment in errors in the process of extinction (5, Fig. 2) means "*exploration*"[3] *in order to find within what region of the already cognitively structured field the goal exists* (or what region has changed into a subgoal). On the other hand, the large number of errors usually found in the beginning stage of acquisition may better be characterized as "*orientation*" *toward structuring cognitively the boundary area between the regions A and G.*

LATENT EXTINCTION IN THE LIGHT OF THE FIELD THEORY

Tolman (12, p. 723) carried out an experiment where rats were trained on a set-up consisting of long and short elevated paths, both leading to food boxes which were behind tall screens. After a strong habit for the short path had been established, the rats were carried by hand to the short-path food box, electrically shocked in this box, and then immediately run in the usual way from the starting place. The results were such that "almost all rats, in spite of the strong habit for the short path, avoided this path, and took the long path." This "latent extinction" or "nonresponse extinction" (2, 9) is one of the serious difficulties the S-R reinforcement theory of extinction faces (3).

Our hodological Equation 5, stated in an earlier article (4), holds also for Tolman's experimental situation, since the direction in the starting place (*A*) to the goal (*G*) along the long path ($w_{A,F,G}$) is *unequal* to that related to the short path ($w_{A,E,G}$), i.e., $d_{A,F,G} \neq d_{A,E,G}$. Latent extinction may, then, be relatively easily explained as due to the changed direction of a psychological force.

DISTINGUISHING BETWEEN LEARNING AND PERFORMANCE

Since our results (4) indicate that even right up to the complete extinction the unity of a path (which corresponds at least *operationally* to ${}_{S}H_{R}$ in terms of the S-R reinforcement theory) as a function of the number of trials[4] was kept at the highest level, it seems to be that extinction has no bearing upon the unity of a response: In terms of Hull's reinforcement theory, there seems to be no relation between habit strength and extinction (4, p. 199). The distinction between *learning*[5] and *performance* which was made originally by Tolman (11, p. 364) may be applied to our case of extinction. If the subject is not motivated, any performance of a learned response may not occur, however perfectly he might have learned the re-

[3] Exploration as such is somewhat similar to what Lewin called "restless movements" which have a direction away from the present region (6, p. 63). Indeed, as often pointed out, subjects remain "frustrated" generally in the process of extinction.

[4] The Equation 8 in the preceding paper (4) may be rewritten, when it is related to the usual conditioning procedure, as

$$uni(w_{CS,CR}) = F(N), \qquad [8a]$$

where *CS* means nothing less than the region which is coordinated to perceiving a conditional stimulus, and *CR* to a conditioned response respectively.

[5] From the field-theoretical standpoint, learning may take two kinds of form: one as shown in the Equations 8 or 8a and the other as shown in the equation

$$stru(LSp) = F(n), \qquad [15]$$

where n indicates the amounts of experience without reinforcement as measured by the definite unit. Of course, Equation 15, which may be viewed as another example of "historical statements" (4, p. 198), does not concern the unity of a path, but the cognitive structuredness of the field, just as in the case of "latent learning" of Blodgett (1).

sponse. This is the case for extinction: Extinction seems to be due to some decrease in the strength of force (drive), that is, D in an equation ${}_sE_R = f(D) \times f({}_sH_R)$. Such a decrease, in turn, accompanies the decrease of potency of the response, in accordance with the Equation 13. Thus, complete extinction of the response ($v_{A,E,G} = 0$) may be viewed as due to $f^*_{A,E,G}$ (in Equation 14) $= 0$. Whether $v_{A,E,G} > 0$ and $v_{A,F,G} = 0$ or $v_{A,F,G} > 0$ and $v_{A,E,G} = 0$ depend exclusively upon the *two-valued direction of the final resultant force* $f^*_{A,G}$ ($d_{A,E,G}$ or $d_{A,F,G}$), and *not* upon the degree of unity of the paths or the cognitive structuredness of the field.

SUMMARY

Both extinction and spontaneous recovery can be viewed as due to the changed direction of a psychological force toward the same goal. Measures such as latency, time required for the response, and errors can be shown not to have the same meaning in acquisition as they have in the process of extinction. One of the principal determinants of the resistance to extinction and of the tendency to spontaneous recovery is the "relative" effortfulness of the response. Both extinction and spontaneous recovery can be viewed as problems of "decision" in an overlapping choice situation. This requires solution of the problem of "potency" of the situation.

REFERENCES

1. BLODGETT, H. C. The effect of the introduction of reward upon the maze performance of rats. *Univer. Calif. Publ. Psychol.*, 1929, **4**, 113–134.
2. DEESE, J. The extinction of a discrimination without performance of the choice response. *J. comp. physiol. Psychol.*, 1951, **44**, 362–366.
3. GLEITMAN, H., NACHMIAS, J., & NEISSER, U. The S-R reinforcement theory of extinction. *Psychol. Rev.*, 1954, **61**, 23–33.
4. JOHDAI, K. Extinction as due to the changed direction of a psychological force. *J. exp. Psychol.*, 1955, **49**, 193–199.
5. JOHDAI, K. The effect of "relative" effortfulness of a response upon extinction and spontaneous recovery. *Jap. J. Psychol.*, 1955, **26**, 87–93.
6. LEWIN, K. The conceptual representation and the measurement of psychological forces. *Contr. psychol. Theor.*, 1938, **1**, No. 4.
7. MILLER, N. E., & DOLLARD, J. *Social learning and imitation*. New Haven: Yale Univer. Press, 1941.
8. MOWRER, O. H., & JONES, H. M. Extinction and behavior variability as functions of effortfulness of task. *J. exp. Psychol.*, 1943, **33**, 369–386.
9. SEWARD, J. P., & LEVY, N. Sign learning as a factor in extinction. *J. exp. Psychol.*, 1949, **39**, 660–668.
10. SOLOMON, R. L. Effort and extinction rate: a confirmation. *J. comp. physiol. Psychol.*, 1948, **41**, 93–101.
11. TOLMAN, E. C. *Purposive behavior in animals and men*. New York: Appleton-Century, 1932.
12. TOLMAN, E. C. A case of learning in rats not explainable by conditioning. *Psychol. Bull.*, 1933, **31**, 723.

(Received September 2, 1955)

Psychological Review
Vol. 63, No. 4, 1956

MEDIATION VARIABLES IN SCIENTIFIC THEORY

WILLIAM W. ROZEBOOM

University of Chicago

It is now nearly a decade since MacCorquodale and Meehl's widely misunderstood paper, "On a Distinction Between Hypothetical Constructs and Intervening Variables" (**9**), appeared in this Journal. Since then, a rash of rebuttals has broken out in the psychological literature, some attempting to redefine the distinction along dimensions other than the one proposed by MacCorquodale and Meehl (e.g., **6**, **10**, **11**), and others repudiating the distinction entirely (**1**, **8**), until a point has been reached where writers on methodological issues no longer treat the two notions as separate concepts, but merely mention in a passing footnote that a distinction may possibly exist.

If this present conceptual chaos were merely a dispute over the naming of formal tools whose usage were nonetheless well understood, the interests of the hypothetical construct (HC)—intervening variable (IV) controversy would best be served by silence. Unfortunately, however, the HC and IV are fundamentally very different in the bases of their introduction into a scientific system, in their confirmation relevant to pre-existent scientific knowledge, and in their potency and functioning as methodological devices. Failure to distinguish between them inevitably leads to the construction of formal models hopelessly at odds with empirical reality, structurally inadequate, inflexible, and presuming precisely the concepts and relations most in need of experimental or semantic clarification. It is advisable, therefore, to re-examine the differences between HC's and IV's in special reference to a number of popular misconceptions that seem to have arisen despite MacCorquodale and Meehl's lucid original analysis.

Introductory Considerations

Preliminary to the major analysis, it is convenient to undertake a terse explication of the notions of "variable" and "empirical law," and a statement of three lemmas in light of which the differences between HC's and IV's will be examined.

Variables. A "scientific variable" is a propositional function—i.e., an incomplete sentence. The kind of variable basic to a scientific system is the "elementary scientific variable." (Other varieties exist, but are either irrelevant for systemic usage or are of a higher logical type.) An elementary scientific variable is an incomplete sentence whose blanks are of two kinds: (*a*) the "reference component," a blank which is filled from a domain of object or event names; and (*b*) the "predicate component," a blank filled from some domain of adjectives or sentence particles such that when the reference component is filled with the name of a specific object, there is one and only one adjective (excluding synonyms) which, when inserted in the predicate component, makes a true sentence.[1] For example, "*o* weighs *x* gms." is a weight variable where *o* ranges over the set of objects under consideration (men, cargoes,

[1] For simplicity, only the case of a single reference and predicate component has been described. More complex elementary scientific variables may have multiple reference and predicate components.

molecules, or whatever) and x covers the domain of positive real numbers. Propositions which are either true or false about an object also fall under this rubric; "o is alive" can be reinterpreted as "It is x that o is alive," which defines an elementary variable when x ranges over the two-valued domain, "true, false." In short, what a variable does is to predicate of each object in the system one of a set of mutually exclusive attributes sharing an underlying similarity which gives the variable its name and distinguishes it from other variables.[2] It is important for conceptual clarity to recognize that although in a technical sense a variable is a relation between an object domain and a predicate domain, the elementary scientific variable is *not* a relation in the sense that the laws of empirical science are relations. Rather, these variables are the units among which such relations hold. Thus to define, with Maze (11), the IV as the empirical relation between observables is not merely to lose sight of the original meaning of "intervening variable" but to contaminate the whole concept of "variable."

Empirical laws. When two or more variables have been defined over the same domain of objects, it is possible that empirical relations exist among them. To say that a (nontrivial) empirical relation exists between two elementary scientific variables x and y[3] is merely to say that although y is not *logically* dependent on x, the value of y for a specific object o can be predicted with greater accuracy when the value of x is known for o than when x is not known. This relation takes the form, $\hat{y} = f(x)$, where $\hat{y}$ is the best estimate of y, given the value of x.[4] Such relations are traditionally referred to as empirical "laws." Unfortunately, the term "law" carries with it connotations of a degree of predictability approaching certainty. In actuality, an empirical law whose accuracy is expressed by a correlation coefficient of $\eta = .01$ has the same systemic and ontological status as a law whose accuracy is $\eta = .99$. It is important to note that an empirical law cannot be discovered unless the related variables have first been explicitly identified. That is, the empirical relations in which a variable participates cannot contribute to the definition of the variable and are thus irrelevant to its ontological content.

I now wish to state three lemmas for future reference. These contain certain cold facts of life which, though hardly controversial, have implications which are usually overlooked by the traditional, highly oversimplified discussion of psychological methodology.

Lemma I. Empirical laws as actually observed by psychologists never attain perfect predictability. We have yet to exhibit an observation variable, psychological or otherwise, which could be predicted with absolute certainty from a set of other known observation variables. This is not a question of whether perfect relations exist in fact, but merely recognition that we, as errant humans, never quite

[2] Formally, ontological similarity of the attribute set defining a variable is superfluous. But even formally the functional similarity remains.

[3] More explicitly, the variables x and y should be written "$P(o, x)$" and "$Q(o, y)$", where P and Q name propositional functions and x and y merely represent the blanks to be filled when a specific value of o is given.

[4] Empirical laws are usually written "$y = f(x)$" with "y" replacing "$\hat{y}$". This has the effect of implying, somehow, that the equation gives the "true" value of y for a given x, and that departure of the real value of y from this computed value is merely a contamination of the law in the same sense that the law of falling bodies is contaminated by air resistance effects (the latter law, of course, being a theoretical derivation and not an empirical relation).

succeed in discovering *all* the empirical determinants of a given variable. Even for empirical laws which seem to admit of no practical uncertainty, we still have a small range of indeterminancy which we ascribe to "errors of measurement."

Corollary. An empirical relation between two observation variables x and y must always be expressed "$\hat{y} = f(x)$" and never "$y = f(x)$." The latter form claims an *exact* relationship between x and y and is hence falsified by the first observation not precisely satisfying the function, no matter how minute the divergence.

Lemma II. No two (logically distinct) observation variables can in practice be demonstrated to have exactly the same determination of a given third observation variable. Because of a popular misconception concerning the notion of predictive equivalences, some explanatory remarks are in order here. It not infrequently occurs that two or more different experimental manipulations have similar effects on the behavior of the organism. Thus we might discover, e.g., that (*a*) a 12-hr. food deprivation, (*b*) a 24-hr. food deprivation with ad lib sugar solution ingestion, and (*c*) a 36-hr. food deprivation with intravenous glucose injection one hour before running all have approximately the same effect on a rat's rate of lever-pressing for food reinforcement. Under such circumstances, we are accustomed to stating that these are alternate variables with the same behavioral consequences. Unfortunately, this terminology, in leaping from operations to variables defined by the operations without explicitly constructing the propositional functions describing the variables, carries the implication that it is possible for only one of a set of variables to hold for an organism. Actually, as commented earlier, each object of a scientific system partakes of *some* value for *every* elementary variable in the system. When we say that two variables x and y are equivalent in their consequences for a third, z, we in reality conceive of x and y having neutral areas in their distributions (e.g., the value corresponding to the rat's *not* having received an intravenous injection), and mean that there are nonneutral values of x conjoined with neutral values of y on the one hand, and nonneutral values of y coupled with neutral values of x on the other hand that have the same predictive consequences for z.

Thus the notion of predictive equivalence among variables is a somewhat misleading abbreviation for the notion that more than one point in a joint multivariate distribution can have the same predictive implication for some dependent variable. That approximations to such equivalence points exist has been repeatedly demonstrated. The claim of this lemma is that we are never able to demonstrate the absolute equivalence of two distinct points. We have many quasi-empirical laws which predict the same value of the dependent variable from multiple points in the joint distribution of the independent variables—for example, all objects having the same ratio of volume to weight have the same prediction as to whether they will float in water. Such laws, however, are never strictly empirical with the prediction from every point independently checked experimentally, but involve the fitting of a simplified equation, usually continuous and frequently involving additional rational assumptions, so that the estimates computed by the equation are seldom strictly the best *empirical* estimates available. This again is not to take a stand on the

factual existence of equivalence points, but merely to point out that we, as ruthless empiricists, are never able to demonstrate them exactly, and for behavioral variables, seldom even approximately.

Lemma III. The empirical laws discovered and formulated by psychologists always presuppose specific fixed values of additional variables. The laws that prescribe the behavior of an organism presume that it is alive and that the neuromuscular structures and environmental supports necessary to mediate the behavior are intact. Again, except for certain special studies, it is further assumed that the animal is not ill, has not received various drug injections, is not deficient in the relevant sensory modalities, and so forth for many other items. That empirical behavioral laws as actually formulated presuppose specific constants in no way detracts from their validity, but it is important to realize that *such laws in no way specify the predictive consequences of their independent variables when the relevant constants are altered.* For a proposition "P implies Q" in no way specifies the implications of not-P.

MEDIATION VARIABLES

Before differentiating between intervening variables and hypothetical constructs, it is expedient to examine the more generic concept of "mediation variable." Let us suppose that $P(o, x)$ and $Q(o, y)$ are propositional functions whose meanings have already been established, and furthermore, that an empirical relationship has been discovered between x and y.[5] This means that given a value for x, we can estimate y as a function of x, a situation which may be symbolized '"x" → "$y(= f(x))$",' the quotation marks indicating that $y = f(x)$ is *not* a law of nature, but rather that $f(x)$ is the value of y which is anticipated by the scientist under belief that x has the given value. To represent this relation without reference to the extrasystemic scientist, we symbolize the relationship, as discussed earlier, by "$\hat{y} = f(x)$." (The latter is able to represent an approximation to a fact of nature because $\hat{y}$ is a property not of a given object, but of the class of objects having a common value of x.)

It is seldom, however, that the scientist—psychologist included—is content merely to infer y from x. Whatever his motives and methodological orientation, he usually avails himself of a mediation variable, $K(o, m)$, the value of which is stated as a function of x and which, in turn, is used to predict y. The scientist's behavior in this respect may be symbolized '"x" → "$m(= g(x))$" → "$y(= \phi(m))$",' recognizing by quotation marks, as before, that the elements related are beliefs of the scientist and not principles of nature. When we attempt to justify the scientist's behavior by translation of this sequence of beliefs into a statement about the properties of the system, however, we find not one possible notation but three: (*a*) "$m = g(x)$, $\hat{y} = \phi(m)$"; (*b*) "$\hat{m} = g(x)$, $y = \phi(m)$"; and (*c*) "$\hat{m} = g(x)$, $\hat{y} = \phi(m)$." In each case, elimination of m can yield no more predictive accuracy than the original empirical relation $\hat{y} = f(x)$, but in case (*a*), m is linked to its antecedent by a rigorous, nonempirical relation, while in (*b*) and (*c*), passage from x to m is empirical and hence subject to error.

In other words, a mediation vari-

[5] For simplicity, we discuss only the case of a single independent variable. The analysis can be extended to the multivariate case by conceiving x as a vector.

able may be introduced as a rigorous function or as an empirical function of its antecedent variable(s),[6] but not as both. If the function is rigorous, the dependency of m on x cannot be empirical (Lemma I) and m must derive analytically from x, with a corresponding analytic relation between the ontological contents of x and m. Conversely, if the dependency of m on x is an empirical relation, the definition and hence the ontological content of m must be independent of that of x, and of the relationship between x and m. Thus investigation of the nature and functioning of mediation variables must proceed along two distinct (though not orthogonal) dimensions: (*a*) the ontological content of the variable—i.e., what it is that the propositional function defining the variable predicates of things; and (*b*) the grounds upon which the scientist is able to justify his introduction of m into the estimation of y from x—i.e., the type of relation between the mediation variable and its antecedents. Much of the confusion which has arisen concerning the differences between HC's and IV's seems to be due to interpretation of the MacCorquodale and Meehl distinction as primarily a distinction of ontological content. Actually, as will become apparent, the ontologies of both HC and IV cover essentially the same range, and it is primarily along the second dimension that the distinction is to be made.

Intervening variables. By "analytic mediation variable," or "analytic variable" for short, let us understand a mediation variable which is derived from its antecedent variables by an analytic relation. Analytic variables, in turn, are of two types: (*a*) primitives, which are logically prior to their antecedent variables, and (*b*) transformation variables, for which logical priority goes to their antecedents. The first type, primitive analytic variables, are of little interest to us here and need be mentioned only for technical accuracy. They comprise the class of mediation variables whose antecedents are logical compounds containing m as a constituent—e.g., the derivation of a person's height given the vector describing his vital statistics, the latter including height as a component. The second type, transformation variables, is our primary focus, for these comprise the class of mediation variables which MacCorquodale and Meehl have named "intervening variables."

Actually, there is little by way of exposition that can be added here to that already presented by MacCorquodale and Meehl. An intervening variable (transformation variable) is a variable which has been systematically *defined* in terms of its antecedents and is hence dependent upon the latter for its meaning. To recapitulate the analysis by means of a formal example, let us suppose that $\hat{z} = (x + y^2)/w$ is an empirical relation expressing the best estimate of a known variable z as a joint function of other known variables w, x, and y. We are now able to express $\hat{z}$ as a function of analytic mediation variables by defining one or more of these as mathematical functions of w, x and y. There are infinitely many different ways in which this can be done.[7]

[6] Here, as henceforth, "antecedent variable(s)" refers to the variable(s) by means of which the mediation variable is systemically introduced.

[7] MacCorquodale and Meehl slightly understated the case when they observed that Hull could have introduced intervening variables into his law of response strength by 15 ways other than the one he in fact chose. The actual number of possibilities is one of the higher transfinities.

To list but a small sample:

$$a =_{\text{def}} (x + y^2), \quad b =_{\text{def}} w^{-1}, \quad \hat{z} = ab; \quad [1]$$

$$c =_{\text{def}} x/w, \quad d =_{\text{def}} y/\sqrt{w}, \quad \hat{z} = c + d^2; \quad [2]$$

$$e =_{\text{def}} 1/2(x + y^2)/w, \quad \hat{z} = 2e. \quad [3]$$

Although an intervening variable is usually introduced so as to simplify as much as possible the written expression of the empirical relation into which it is introduced, *any* transformation of the antecedent variables into an arbitrary domain of symbols is a perfectly legitimate definition of an IV, and can (with the perhaps necessary assistance of other similarly defined IV's) replace its defining variables in the empirical equation under consideration. Intervening variables are of frequent and important application in statistical analyses when it is desired to transform a distribution of raw scores into a distribution, such as the unit normal distribution, with special mathematical properties.

In description, then, of the nature of the IV along the second dimension mentioned above, we see that the IV is derived nonempirically from its antecedent variables. That is, passage from the latter to the former does not admit of uncertainty, for when an object's values of the antecedent variables are known, its value of the IV is also rigorously defined. On the other hand, the empirical law relating the IV to the dependent variable can at best be no more accurate than the original unmediated empirical relation and actually represents a loss of predictive accuracy unless the definition of the IV in terms of its antecedents is a one-one function. This is because in general, as pointed out in Lemma II, no two points in a multivariate distribution ever have quite the same strictly empirical prediction to the dependent variable, and whenever two or more such points define the same value for the IV (as obtains, for example, when the IV is defined as a ratio of two antecedent variables), prediction from that value of the IV never can have quite the same predictive accuracy as prediction from any one of the corresponding values of the antecedent variables. As for the ontological content of the IV, it is precisely the same as that of the antecedent variables which define it. There are various technical devices by which a definition can be accomplished in a formal system—explicit definition, recursive definition, reduction sentences, etc.—but all have the common characteristic that any sentence containing a defined term can be replaced by a sentence not containing that term. If $K(o, m)$ is an IV defined in terms of an antecedent variable $P(o, x)$, and $a_1, a_2, \cdots, a_n$ are specific values of x generating the same specific value, b, of m, then the assertion "$K(o, b)$" about an object o has the same meaning as "$P(o, a_1)$ or $P(o, a_2)$ or $\cdots P(o, a_n)$." To say "o has a density of 2 gm./cc." is to say, "Either o has a volume of 2 cc. and weighs one gm., or has a volume of 3 cc. and weighs 1.5 gm., or $\cdots$, or $\cdots$" etc. The latter sentence completely exhausts the meaning of "density."

Hypothetical constructs. In the last section we considered those mediation variables which derive analytically from their antecedents. For the remainder, passage from antecedents to mediators must be made with less than logical certainty, and such variables we shall designate as "inferred variables." Precisely because inferred variables are not analytically contained in their antecedents, it follows immediately that (*a*) the meaning, or ontological content, of the inferred

variable is *not* reducible to that of its antecedents, and (*b*) a certain degree of error must be expected in passage to the inferred variable from its antecedents.

With some reluctance, we now distinguish between two main types of inferred variables: (*a*) inferences where the specific value of an object is estimated on a variable already accepted as existent for that object, and (*b*) inferences where not only specific values are estimated, but where for that object the very existence of the generic property defining the variable is subject to critical doubt.[8] As a practical example of the first type, we might conceive of a situation in which a construction foreman, needing a heavy worker to ballast an unriveted beam, selects the tallest and widest man he can find. Although the foreman has probably never done an empirical study relating height and width to successful ballasting, he knows that on the one hand, ballasting ability is directly determined by weight, and that on the other hand, weight in humans can be closely inferred from height and width. To illustrate the second type, we may take the traditional example of the kinetic theory of gases and the empirical gas laws. Given two of the three observation variables—pressure, temperature, and volume—for an enclosed volume of gas, we can infer the distribution and kinetic energy of the gas particles, and deduce the third observation variable, in turn, from this inference. In addition, however, we must hypothesize that the property of being composed of a large number of small elastic particles does, in fact, exist and hold for gases.

Because the very existence of inferred mediation variables of the second type is inferential, and because, as seen below, such properties can be creatively imagined, the term "hypothetical construct" is singularly appropriate for variables of this sort. However, this term calls undue attention to the unverified nature of the variables, and not infrequently leaves the impression that such variables are flights of metaphysical fancy. Actually, all properly formulated HC's, even the more abstract ones, maintain close contact with testable reality and many involve only the most pedestrian kind of speculation. In fact, it can be argued (and argued convincingly, in my opinion) that *all* scientific variables are hypothetical constructs or transformations thereof, even such primitive observables as length, weight, lever-pressing frequency, etc., for even these, as part of their implicit definitions, contain notions of properties such as constancy under spatial translation, existence in a "real" world, etc. which far transcend the immediately given experiences of the scientist. Since MacCorquodale and Meehl have stressed the existential, rather than the hypothetical, aspects of HC's, I suggest that by "hypothetical constructs" we understand all cases of inferred mediation variables, even though in a few instances, the generic existence of the variable and its applicability to the objects concerned may admit of little practical doubt.

At this point, it becomes mandatory to say a few words about the ontological content of hypothetical constructs, even though philosophers have yet to produce any substantial agreement on this (cf. **5** for a comprehensive summary of the major inter-

[8] To claim, prior to knowledge of the specific value of x for o, that a variable $P(o, x)$ holds for an object o is to assert the proposition "$(\exists x)P(o, x)$." The present distinction is in reference to whether or not this proposition is considered doubtable.

pretations). There would appear to be three distinct levels of uncertainty for the existence of a HC. On the lowest level are variables whose generic defining property is already accepted as existent, but whose legitimate predication of the objects being dealt with is uncertain.[9] Thus in regard to the variable, "The particles comprising the gas in container o have a total kinetic energy of x," we already have a positivistic understanding of an aggregate of discrete particles—as, for example, experientially gleaned from observing a table of pool balls—and the properties abstracted from their motions, but do not know whether, in fact, the gas in container o is comprised of such particles. Inferences on this level involve no more methodological subtlety than does inferring from a man's stated preferences in female beauty the color of his wife's hair, even though it may not be known whether the man is married.

On the second level of uncertainty are those hypothetical constructs consisting of a compound predicate whose constituent properties are accepted as existing separately, but whose conjoint existence is uncertain. Most of the hypotheses of everyday life are on this level. Thus as Bergmann (1) points out, to speculate on the existence of a unicorn is simply to hypothesize that the properties describing the appearance of a horse and the property of having a long horn on the forehead exist conjoined in a single object. Or when we infer from an empty rat cage and scurrying sounds under the cage rack that an escaped rat is in the corner of the laboratory, we hypothesize that there is an object which possesses both the properties of being small, white, furry, etc., and being located at the spatial-temporal coordinates of that corner.

Third-level constructs, a more sophisticated version of level two, and the level comprising the bulk of scientific constructs, are also compounded of separately defined predicates. On this level, however, at least one member of the compound is a relational predicate entailing a functional association, usually a rigorous or semi-rigorous[10] one between the HC and other variables of the system. Thus part of the hypotheses underlying the kinetic theory of gases is that the particles of which a gas is presumably comprised obey certain laws of elasticity, conservation of momentum and energy, etc. which permit the derivation of the observable properties of gases as a rigorous analytic function of the underlying HC's (subject to certain secondary disturbances). We shall return to the relational assumptions of the HC in slightly greater detail later.

It is important to note that the HC is no more unverifiable than the predicates from which it is constructed. For constructs of the first and second level, whatever test procedures are considered adequate to verify the constituent predicates taken separately are in principle jointly applicable as a test for the HC. Thus the particulate nature of gases is in principle directly verifiable (or refutable), given a sufficiently powerful magnification device. The relational predicates of the third level are not so easily tested directly, but since such predicates need be built into the

[9] Formally, this level comprises variables for which "$(\exists o)(\exists x)P(o, x)$" is accepted, but where for a specific object o, "$(\exists x)P(o, x)$" is questionable. For second and third level constructs, even "$(\exists o)(\exists x)P(o, x)$" is open to practical doubt.

[10] By saying that the determination of y by m is "semi-rigorous," we mean that m exactly determines y by the specified relation so long as certain supplementary variables remain constant or under certain theoretical constraints.

construct only when some testable covariation among observables is entailed, they are subject to techniques for inductive corroboration.

The feature of the HC which gives its particular fruitfulness as a mediation variable and its convincingness as an explanatory device is precisely that feature which has caused some of our more methodologically chaste psychologists to protest in moral indignation—the "excess meaning" of the HC. Since the HC is defined independently of the variables from which it is inferred, the HC may have analytic implications or known empirical relations far beyond anything known or surmised about its antecedents. In regard to first- or second-level elements of the HC, we may have much independent empirical knowledge concerning the behavior of constituent properties. Suppose, for example, that, in the analysis of visual discrimination, we hypothesize that discrimination behavior is mediated by electromagnetic excitation of retinal cells. By this inference, we have immediately brought into consideration a mass of empirical data concerning electromagnetic waves and retinal cells arrived at independently of discrimination research (e.g., the observation that electromagnetic waves will not pass through certain types of obstructions or that retinal cells have such-and-such limits of response) which would otherwise be irrelevant to discrimination problems but now permits prediction of the dependent discrimination variable (e.g., prediction that the organism cannot discriminate wearing opaque goggles, or that discrimination cannot exceed such-and-such acuity). With third-level constructs, the relation of the dependent observation variable to the HC is either rigorously or semi-rigorously entailed by the constitution of the HC. When we have an empirical relation $\hat{y} = f(x)$, and insert a HC as a mediation variable so that $\hat{m} = g(x)$, $\hat{y} = \phi(m)$, the uncertainty of passage from m to y may be as little as we please, depending upon our construction of m. Passage from x to m must be a calculated risk, not merely in inference to a specific value of m but to the very existence of m. Once the risk is accepted, however, we have not only a "rational" explanation for $\hat{y} = f(x)$, but also (*a*) prediction as to the relation of x and y beyond the actual limits tested empirically at the time, and (*b*) predictions of empirical relations between x or y and other variables designated through their relations to m but bearing no intuitive relevance to the $\hat{y} = f(x)$ association. That is to say, inference to the existence of an HC inevitably entails many more predictions to empirical relationships than merely the one for which the construct was initially invoked.

Let us now consider some attempts to clarify several misconceptions that seem to have arisen concerning intervening variables and hypothetical constructs.

Intervening variables do not lack existential reference. MacCorquodale and Meehl have been accused by Ginsberg (**6**) of defining the IV as something entirely arbitrary and without existential reference. We have just seen, however, that while the algorithmic device by which we formally define and compute the IV is essentially arbitrary (as are, indeed, all our symbolic techniques), the ontological referent of the IV is the same as that of the variables from which it is derived. Hence if the antecedent variables designate "real" properties, so must the IV, and similarly, the latter must be empirically

related to any other variables that are empirical functions of its antecedents.

The hypothetical construct is not a "convenient fiction." There seems to be a widespread belief, not merely among psychologists, that HC's are "convenient fictions," or artificial formal devices. If the properties which are hypothesized to co-exist in the HC have "real" referents individually, they do not somehow become fictional just because their conjoint existence has not as yet been verified. An HC cannot simultaneously be both convenient and fictional, because if the hypothesis expressing the HC is in fact false, the entire system incorporating the HC becomes basically useless. For a conjunction of propositions is falsified by the falsity of any one of its constituents.

The hypothetical construct is not unobservable in principle. Unless we adopt an extreme solipsism and assert that, e.g., the objective world exists only when we see it, there is nothing about an HC which in principle bars it from direct confirmation. This has already been discussed above, and the point is simply illustrated through the example of a mechanic diagnosing a malfunctioning automobile. His inference of dirt in the carburetor is to a hypothetical construct, but to verify or reject the hypothesis, all he need do is to look in the appropriate part of the engine.

The excess meaning of the hypothetical construct does not consist of vague, undesirable prescientific connotations. I suspect that much of the distrust shown toward HC's by some of our more tough-minded psychologists has stemmed from fear that the HC is a methodological Trojan Horse by which mentalistic notions such as "idea," "desire," "self," etc. are surreptitiously to be re-introduced into behavioral psychology. Any vagueness inhering in an HC, however, can only be due to vagueness in the definitions of its component properties. It is true that the ramifications of an HC may not at first be fully comprehended, but these additional commitments to testable expectancies are precisely wherein the potency of a good HC lies. Vagueness in a concept, on the other hand, has just the opposite systemic effect. The reason that older mentalistic notions or more modern refinements such as "libido," "r_g," etc., have proven objectionable is precisely because their lack of explicit description prevents derivation of testable, unambiguous consequences.

The intervening variable is not a manifestation of, nor is the hypothetical construct the antithesis of, methodological rigor. Since an IV is merely another way of expressing its antecedent variables, its introduction in no way adds to or subtracts from whatever formulational rigor the system may possess. On the other hand, while it is not difficult to use HC's poorly, their proper employment necessitates explicit delineation of essential properties and careful derivation of both empirical and analytic implications to testable propositions—demanding, in short, a high degree of methodological rigor.

Intervening variables and hypothetical constructs do not represent different degrees of operational validity. Marx (**10**) has attempted to redefine IV's and HC's as distinguishable only as different degrees of operational validity, high for the former and low for the latter. I am no longer certain what "operationally valid" means, since the term has all but lost its operational definition and has, rather, become a synonym for "good concept" or an antonym for "sinful." The notion, however, scarcely seems ap-

plicable to the IV–HC distinction. Since the IV is analytically derived from its antecedents, the former has precisely the same operational validity as the latter, which may be of any degree. Similarly, the range of operational validities for HC's must cover almost the entire spectrum of possibilities, according to the uncertainty of the inference by which the HC is introduced.

Neither the intervening variable nor the hypothetical construct can be "anchored on both sides." The requirement that a mediation variable be anchored on both sides, first introduced by Hull (**8**, p. 22), stipulates that a mediation variable must be introduced from observation variables by an equation or chain of equations, and must, in turn, lead by an equation or chain thereof to another observation variable. The fact that all empirical laws actually known to psychologists are but imperfect estimates of the dependent variable precludes the linkage of dependent and independent variables by any chain of exact equations. This means that in any chain of mediation variables linking observables, at least one link must be nonanalytic and hence irrelevant to the definition of the mediation variables. Thus a mediation variable bracketed by observation variables can be anchored rigorously to at most one end of the empirical relation mediated, while the remaining linkage remains strictly empirical and hence logically unnecessary.

Laws containing only intervening variables have no inductive potency. The mathematical expression for a finite number of points may be extrapolated by an infinite number of alternate equations. Also, as seen earlier (Lemma III), an empirical law always presupposes specific fixed values for other implicit variables. Hence an empirical law may be considered reliable only for the actual values of the variables tested, and provides no justifiable predictions for the relation of the empirical variables outside the range tested, much less to predictions of the relation when the values of the implicit variables have been changed. Since a law containing only IV's is merely shorthand for an empirical relation, we must conclude that such laws cannot rationally be extrapolated beyond the data from which they were derived without additional assumptions. On the other hand, if passage from the HC to the observation variables is in part analytic (as is almost always the case), or if the excess meaning of the HC permits an evaluation of its relation to the implicit variables held constant in the original empirical law, then introduction of an HC as a mediation variable into an empirical relation may extend the relation (subject to experimental disconfirmation) far beyond the limits of the original observations.

The intervening variable is NOT a causal mediation variable. Our earlier analysis of IV's as a type of mediation variable employed the term "mediation variable" strictly to designate a syntactic property—i.e., to indicate a formal expression written by the scientist between his formal symbols for the dependent and independent variables of an empirical relation. We frequently have reason to believe, however, that given an empirical relation in which a variable y covaries with a variable x, there exist one or more "real" variables v whose identities may be unknown to us but which causally mediate between x and y. That is, we suspect x tends to determine y only because x tends to determine v and v, in turn, causally influences y, so that if v were to be

prevented from varying, the relation between x and y would vanish—e.g., as in the case of the relation between the first and last gear in a train of gears. Now, since an HC has an ontological content independent of that of its bracketing variables, it can and frequently does occur that the HC, if it in fact exists, behaves as a causal mediator for the empirical relation which it is invoked to explain. Thus, neurophysiological constructs such as nerve impulses, etc., are frequently invoked as causal mediators between stimulus and response variables. Intervening variables, on the other hand, can in no way represent causal mediators, since the IV is merely a syntactic transformation of its antecedents and has the same ontological referent as the latter. Similarly, the breaking up of a complex empirical relation into several simpler equations through the utilization of several intervening variables in no way need reflect the dependency structure of causal mediators, since the resolution into IV's is entirely arbitrary.

The intervening variable does NOT tie together sets of independent and dependent observation variables and permit a reduction in the number of laws describing their interrelations. As originally observed by Feigl (**4**) and recently spelled out in detail by Seward (**12**), a mediation variable may permit a sharp reduction in the number of laws required to express the relationship between k independent and n dependent variables. Taken pairwise, k independent and n dependent observation variables determine kn empirical relations. If these relations follow a certain pattern, however, a common mediation variable m can be written as separate functions of the independent variables—$m = f_1(x_1), \cdots, m = f_k(x_k)$—and the dependent variables, in turn, written as individual functions of the mediator—$y_1 = \phi_1(m), \cdots, y_n = \phi_n(m)$—so that the total number of equations becomes reduced to $k + n$. Seward has explicitly claimed to find no distinction between HC's and IV's in regard to such a mediator.

The discerning reader will observe immediately that the above system of interrelations, where k independent variables appear to be equivalent in the determination of each of n dependent variables, is a more complex form of the situation previously discussed in Lemma II, concerning the equivalence of more than one variable in the determination of a single dependent variable. It now becomes apparent that a mediation variable which functions in the manner just described cannot be an IV but must, rather, consist of a HC. Reasons:

1. An IV cannot be defined simultaneously by two or more independent functions. To define, e.g., an IV simultaneously as $m =_{def} f_1(x_1)$ and $m =_{def} f_2(x_2)$ implies that $x_2 = f_2^{-1}f_1(x_1)$, thus asserting a rigorous covariation between x_1 and x_2 which in practice never occurs when x_1 and x_2 are not analytically equivalent. Hence defining an IV in terms of more than one observation variable can only be accomplished by: $m =_{def} [f_1(x_1)$ or $f_2(x_2)]$. But since observation variables x_1 and x_2 are never demonstratively related in the same way to the dependent observation variable y, prediction of y from m is less accurate than from either x_1 or x_2, and the value of y predicted from m differs from that predicted from either x_1 or x_2. In other words, since no exact equivalences can be found among k independent observation variables in their determination of n dependent variables, it is not mathematically possible exactly to reproduce the kn empirical relations by $k + n$ equations with an IV as the mediation variable. For psychological laws in particular, where we have yet to demonstrate even close equivalences, the loss of predictive accuracy through attempting to introduce a single IV as a common mediator would be tremendous. On the other hand, when k independent variables seem reasonably similar in their determination of n dependent variables, it seems very reason-

able to suspect the existence of some additional variable, a HC prior to its actual detection, whose value is substantially determined alternatively by x_1 or x_2 or ... or x_k and which, in turn, substantially determine $y_1, y_2, \ldots,$ and y_n. Thus for the example in Lemma II, where different nutritional schedules were presumed to have similar effects in determining lever-pressing rate, it would seem not illegitimate to conclude that while these operations are not identical in their influence on lever pressing and might, for example, be quite different in their cue value (S^D) for that behavior, they may nonetheless similarly influence an underlying behavioral variable called "degree of hunger" whose meaning is independent of the operations by which it is aroused.

2. A mediation variable is frequently inserted into a matrix of k independent and n dependent variables before all kn empirical relations have been experimentally determined. Thus Seward (12) speaks of testing the validity of a mediator by observing empirical relations: $\hat{y}_1 = g_1(x_1)$, $\hat{y}_2 = g_2(x_2)$; inserting a mediation variable: $m = f_1(x_1)$, $\hat{y}_1 = \phi_1(m)$; $m = f_2(x_2)$, $\hat{y}_2 = \phi_2(m)$; and *subsequently* testing the validity of the inferred relations: $\hat{y}_2 = \phi_2(f_1(x_1))$ and $\hat{y}_1 = \phi_1(f_2(x_2))$. When m is an HC, inferences to previously untested empirical relations may easily be justified by the excess meaning of the construct. On the other hand, if such a mediator is an IV, there is no reason to infer that the complete $k \times n$ covariance structure is of the sort reducible to $k + n$ laws. For no IV can provide a basis for any inference not already inferrable directly from the empirical relations into which the IV has been introduced.

THE CONTENT OF HYPOTHETICAL CONSTRUCTS

By now, I hope, the reader will have come to regard with some favor the writer's view that the methodological properties of the intervening variable and hypothetical construct are drastically distinct; that the IV, while aseptically respectable, achieves this through inductive, explanatory, and hueristic sterility; that the HC, precisely because of its fertility, involves a calculated inferential risk and a certain virtuosity in construction; and finally, that either to deny the legitimate usage of the latter or to introduce the latter with the complacency entitled by the former is to demonstrate a less than adequate comprehension of basic scientific methodology. From this, it follows immediately that one of the most urgent problems facing psychological theory is an explication of the more detailed nature of hypothetical constructs and of their optimal construction and employment. While such an undertaking is considerably more formidable than any possible of attempt here, I wish, nonetheless, to offer some suggestions as to the necessary and sufficient ontological content of HC's suitable for behavioral theory.

It has been almost unanimously presumed that any HC invoked to account for molar behavioral principles must constitute an inference specifically about neurophysiological processes underlying that behavior. This belief, coupled with the faith that behavioral laws are capable of formulation on a molar level alone, has doubtless been a major factor in the aversion for HC's shown by some behaviorists, and why others have attempted, through redefinition, to infuse the term "intervening variable" with the formal properties of the hypothetical construct (cf. **2**, for an especially flagrant example). Personally I have the fullest sympathy for those who patiently seek an explanation of organic behavior in terms of neurophysiological mechanisms. That molar behavioristics is in theory completely reducible to underlying neurophysiological principles is not a hypothesis but a logical derivation from the fact that every instance of a response is analytically concomitant with a specific set of neurophysiological events, so that a completely deterministic neurophysiology must of necessity permit derivation of

all molar behavioral laws. It is seriously to be doubted, however, whether the present state of physiological knowledge is sufficient for an HC constructed of neurophysiological predicates to have many logical or empirical implications for behavioral relations. Most, if not all, studies permitting an integration of behavioral and neurophysiological principles have resulted in extension of our knowledge not so much about behavior as about neurophysiology. For while existent neurophysiological data impose little or no restrictions on possible behavioral relations, existent knowledge of the principles of behavior prescribes an important set of minimal properties that neurophysiological processes must possess.

Neurophysiological content is not only at present inadequate for behavioral HC's, but also basically irrelevant. Presuming the switchboard theory of neural action to be true, we could in principle replace each neurone with a mechanical or electronic device with the same connections and input–output characteristics as the replaced cell without affecting the validity of any behavioral law. Since neurophysiological attributes are thus not essential for a physical behavior system, a set of HC's need possess no neurophysiological content in order to describe the necessary properties of the causal mediators underlying behavior. Nor does the number of HC's required for an adequate theory of behavior need be on the order of the number of neurones in the nervous system any more than the kinetic theory of gases requires a separate variable for each gas molecule. The functional units of behavior—i.e., the essential properties of the system which actually make a difference for behavioral outcomes—are undoubtedly properties of neural aggregates or mass actions, just as in a Hebbian brain the functional unit is the phase sequence, the behavior of which can be formulated without reference to any constituent neurones.

I suggest, therefore, (*a*) that despite the complexity of the central nervous system, those of its properties which are essentially involved in the causal determination of the dependent variables of behavioral psychology may possibly be reflected by a relatively small number of HC's without neurophysiological reference; and (*b*) that the necessary ontological content of such HC's need consist only of existential quantification (the logical operator "$\exists x$") and relational predicates. In conjunction with this latter suggestion, I wish to point out that an HC may be introduced with no more meat on its bones than "There exists a variable m such that for observation variables x and y, $\hat{m} = g(x)$ and $\hat{y} = \phi(m)$." For example, given the existence of an empirical relation $\hat{z} = wxy$, we might hypothesize, "There exist variables m and n such that $\hat{m} = wx$, $\hat{n} = y$, and $\hat{z} = mn$." It is very important to distinguish such a hypothesis from a definition of intervening variables: "$m =_{\text{def}} wx$, $n =_{\text{def}} y$, and $\hat{z} = mn$." The existential hypothesis asserts an *empirical* relationship between the mediation variables and their antecedents, thereby leaving room for (*a*) additional hypotheses concerning the nature of m and n, (*b*) possible independent substantiation of the existence of m and n, and (*c*) improvement or correction of the inferred relations among the mediation and bracketing variables in the light of advancing knowledge.

But most important of all are the *structural* inferences contained in the existential hypotheses defining HC's.[12]

[12] This point seems to have been anticipated by MacCorquodale and Meehl. "A hypo-

Decomposition of a complex empirical relation by means of IV's has no necessary relation to the covariance structure of the causal mediators underlying the empirical relation. An intervening variable breakdown, "$m =_{\text{def}} wx$, $n =_{\text{def}} y$, $\hat{z} = mn$," has precisely the same ontological meaning and predictive implications as "$m =_{\text{def}} w$, $n =_{\text{def}} xy$, $\hat{z} = mn$." To hypothesize "$\hat{m} = wx, \hat{n} = y, \hat{z} = mn$," on the other hand, involves expectancies, e.g., about the manner in which the variables hang together under altered supporting conditions, which differ from the expectancies involved in the hypothesis "$\hat{m} = w$, $\hat{n} = xy$, $\hat{z} = mn$" and which are capable of experimental substantiation. Thus if we were to discover that under the influence of a certain drug, the empirical relation in the present example became changed to $\hat{z} = y\sqrt{wx}$, we can readily interpret the result by inferring that $\hat{m} = wx, \hat{n} = y$ and that the drug changes the manner by which m enters into the determination of z to $\hat{z} = n\sqrt{m}$. This is a more parsimonious interpretation of the drug result than any which can be made under the hypothesis that $\hat{m} = w$, $\hat{n} = xy$.

Once we realize, in fact, that the structure of a system of HC's is in itself a predicate sufficient to carry a set of existential operators, we become aware that it is not even necessary for the hypothesis to specify in quantitative detail the functional relations by which the variables are connected. It is sufficient that the hypothesis merely designates for each construct the other variables of which it is an immediate function and those which, in turn, the construct immediately helps determine. Thus if we treat the mediation variables of Hull's system not as IV's but as HC's (which, of course, the excess meaning introjected by Hull reveals them to be), we are, if we wish, able to conceive D, ${}_{S}H_{R}$, etc. as real attributes of the organism, not necessarily burdened with specific neurophysiological content, whose exact quantitative relations to each other and to extra-organismic variables still await experimental clarification, and yet variables whose structural assumptions generate testable predictions, especially when determination of the more detailed nature of these interrelations is regarded as a major experimental objective. Recognition that the structure of causal linkages among a set of mediation variables is independent of the specific quantitative nature of the relations frees us from preoccupation with excessive quantification of empirical relations during the early phases of a science, enabling us to identify the major immanent determinants of behavior prior to discovery of their exact relations to their causal antecedents and observable consequences.

thetical construct which seems inherently metaphorical may involve a set of properties to which hitherto undiscovered characteristics of the nervous system correspond. So long as the propositions about the construct are not stated in the terms of the next lower discipline, it is always a possibility that the purely formal or relational content of the construct will find an isomorphism in such characteristics. For scientific theories this is enough, since here, as in physics, the associated mechanical imagery of the theorist is irrelevant" (**9**, p. 106; cf. also "The Nomological Net," **3**, p. 290).

REFERENCES

1. BERGMANN, G. The logic of psychological concepts. *Phil. Sci.*, 1951, **18**, 93–110.
2. BERGMANN, G. Theoretical psychology. *Annu. Rev. Psychol.*, 1953, **4**, 435–458.
3. CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, **52**, 281–302.

4. FEIGL, H. Operationism and scientific method. *Psychol. Rev.*, 1945, **52**, 250–259.
5. FEIGL, H. Existential hypotheses. *Phil. Sci.*, 1950, **17**, 35–62.
6. GINSBERG, A. Hypothetical constructs and intervening variables. *Psychol. Rev.*, 1954, **61**, 119–131.
7. HULL, C. L. *Principles of behavior.* New York: D. Appleton-Century, 1943.
8. KENDLER, H. H. "What is learned?"—A theoretical blind alley. *Psychol. Rev.*, 1952, **59**, 269–277.
9. MACCORQUODALE, K., & MEEHL, P. E. On a distinction between hypothetical constructs and intervening variables. *Psychol. Rev.*, 1948, **55**, 95–107.
10. MARX, M. H. Intervening variable or hypothetical construct? *Psychol. Rev.*, 1951, **58**, 235–247.
11. MAZE, J. R. Do intervening variables intervene? *Psychol. Rev.*, 1954, **61**, 226–234.
12. SEWARD, J. P. The constancy of the *I–V:* a critique of intervening variables. *Psychol. Rev.*, 1955, **62**, 155–168.

Psychological Review
Vol. 63, No. 4, 1956

THE CONTRIBUTION OF JOHN B. WATSON [1]

GUSTAV BERGMANN

State University of Iowa

Second only to Freud, though at a rather great distance, John B. Watson is, in my judgment, the most important figure in the history of psychological thought during the first half of the century. Nor is his impact limited to the science of psychology. Understood or misunderstood, quoted or misquoted, his name and his work are a symbol around which debate has swirled for quite some time, not only among psychologists but also generally among intellectuals of our civilization, although the attention he receives now, in the fifties, is perhaps not as great as it was in the twenties and thirties. There are two reasons for this eclipse, which I believe is temporary. Among psychologists the sound core of Watson's contribution has been widely accepted; his errors and mistakes have been forgotten. Naturally, then, psychologists have less reason to discuss either. In our civilization at large, the center of the debate has, during the last ten or fifteen years, shifted so far toward the antiscientific—and often quite irrational—right that Watson's bitterest enemies, who have always quoted him most eagerly, feel that they need no longer bother to refute him. They only use his name occasionally to scare little children in the existentialist dark. For both these reasons Watson, though happily alive, is by now a historical figure. So it is perhaps not too rash if one tries to look at both the man and his work historically.

Most of what I shall say will be about the work. Let us then first briefly glance at the man. There is a French saying that the style is the man. If it is right, as I believe it is, then one of the best ways to get the flavor of Watson's intellectual personality is to turn to his writings. Doing that, one receives above all the impression of simplicity, lucidity, vigor, and intellectual courage. Reading on, one is soon struck by the vices which in this case go with these virtues. The simple at times shades into the *simpliste;* the intellectual strength is often without subtlety; even the courage becomes less attractive when blended with the recklessness of the extremist. These are strong words, both for better and for worse. If they are justified, then there is much light as well as much shadow. Such men make a certain kind of contribution. So I shall inquire whether the analysis of the contribution agrees with the thumbnail sketch of the man. I believe that it does.

Watson is not only an experimental *psychologist,* fairly distinguished no doubt but without transcendent distinction; he is also a systematic thinker, that is, a philosopher of psychology, or, as one says, a *methodologist.* In this latter area he made his major contribution. But he is not only a methodologist, he is also a *social philosopher*—I am almost tempted to call him a social prophet—and he is, finally, a *metaphysician.* Since he is, vigorously and unsubtly as usual, a man of one piece, these four sides of him—the psychologist, the methodologist, the prophet, and

[1] This is, with very minor alterations, the text of an address delivered before the Psychology Club of the University of Chicago and at Northwestern University. I have tried to preserve as much of the livelier and less formal tone of the spoken word as was possible and seemed proper; and I have, in the same vein, dispensed with the apparatus of references and footnotes.

the metaphysician—have often interfered with each other, very much to the *logical* detriment of each.

Let me set the tone of this talk by anticipating some of the judgments at which I shall eventually arrive. I shall be quite blunt—as blunt, perhaps, as Watson himself would be. His social philosophy is, in my opinion, deplorable. His metaphysics is silly. (Philosophers use this word, silly, without malice; e.g., a realist might say that phenomenalism is a silly philosophy and yet recognize that some of the arguments made in its favor are ingenious or even, as he thinks, irrefutable.) But then, again, things are not that simple; they are, Watson to the contrary notwithstanding, neither all white nor all black. Thus I hasten to add that the rationale behind his social gospel was a very high-minded, though naive and, alas, misguided type of idealism fairly widespread in this country at the beginning of the century. His metaphysics, quite similarly, has a core both true and important. The trouble there was that Watson, unsubtle and, I fear, also very ignorant in philosophical matters, did not know how to state the true and commonsensical core without at the same time asserting a lot of patent nonsense. ('Nonsense' is another word philosophers use without malice.) But, again, it is only fair to add that in this he was no worse, though he certainly was no better, than most of his fellow metaphysicians—technical philosophers and philosophical scientists alike.

Leaving the social philosophy to the end, I begin with an examination of the metaphysics. It should help if I anticipate the result of this examination. The thesis Watson was above all eager to establish and, man of action that he was, to inculcate, is that *there are no interacting minds*. This thesis is true beyond reasonable doubt. It is in fact sheer scientific common sense. Thus it is metaphysical only in the peculiar sense that, as experience has taught us, not only in the case of Watson, the philosophical analyst's professional skill is required to state it correctly, that is to say, without at the same time asserting a lot of patent nonsense. Watson's particular mistake was that in order to establish that there are no interacting minds, which is true, he thought it necessary to assert that *there are no minds*, which is not only false but silly. This, then, is the pattern. Watson's conclusion, that there are no interacting minds, is true and, certainly by now, sheer common sense. His premise, that there are no minds, is false and sheer nonsense. But I notice that at this point some explanations are in order.

First, with respect to minds—or, rather, the word 'mind'—I do not use it for that ancient notion which is traditionally called a substantial mind. Nor would it make sense historically had Watson tried to lay once more this particular ghost; substantial minds had been banned effectively from psychology by Wundt a generation earlier. What I mean, therefore, when for brevity's sake I speak of minds, are so-called mental contents, or awarenesses, or phenomenal givennesses, such as percepts, volitions, memories, thoughts, and so on. That there are such things and that they are not, literally, physical things, either inside or outside the human body, is obvious. It is not the task of logical analysis either to prove or to refute the obvious. The task is, rather, to keep us out of trouble when talking about it. So I turn next to an explanation of the word 'interaction' and of the phrase 'interacting minds.' I shall make use of what philosophers of science call a thought experiment.

Imagine a world with space and time like ours but otherwise so impoverished that it contains only Newtonian mass points which are colored, each of them

being either red or green or blue. The mass points obey Newton's law; that's why I call them Newtonian. This law is a process law, which means that, given the masses, the positions, and the velocities of a (closed) configuration at a moment, one can, by means of this law, compute (predict) their positions and velocities at any earlier or later moment. In this respect my imaginary world is thus like ours. With respect to color it is very different. The whole of its space is divided into three regions, call them *R*, *G*, and *B*, respectively, such that whenever a mass point is in *R* it is red, whenever it is in *G* it is green, and so on. This world thus has two fundamental laws; one is Newton's law, which is of the process type; the other is the color law I just stated, of the type called cross-sectional. The point of the story is, of course, that it permits me to explain the relevant meanings of 'interaction' and of the related phrase 'causally closed.' Using them correctly in speaking about this world of my fancy, one would say that there are colors in it but that these are noninteracting colors. Or one might say that this world minus its colors is causally closed under the Newtonian process law. But, I repeat, to say this is not to say either that there are in this world no colors or that in it there does not hold a law, of the cross-sectional variety, which coordinates to each position a color. The moral of the fable is obvious. To say that there are no interacting minds is to say, first, that men's bodies with all their stuffings, including of course their central nervous systems, are part of the physical universe, and, second, that the physical universe is causally closed under laws none of which ever mentions anything mental. But, to repeat once more, to say this is not to say that there are no minds.

Anyone with some knowledge of the intertwined histories of psychology and philosophy may at this point wonder why I have bothered to elaborate the obvious. For you may remember, from the history of philosophy, that the thesis I have just explained so meticulously was, at Watson's time and before, well known and widely accepted under the name of epiphenomenalism. I shall counter the question with another. Why, I ask, did Watson himself bother? Why didn't he simply accept some version of epiphenomenalism, particularly since it was, at least latently, also the philosophy of the Functionalists, and let it go at that? Nor shall I stop there but go on to point out that, as you may remember from the history of psychology, Watson, as he saw himself, rebelled not only against Wundtianism, or, as it was called in this country, Structuralism, but also and at least as vigorously against the Functionalism of his teachers at the University of Chicago. This is prima facie evidence for something more than meets the eye at a first glance. It is also my cue to say what needs to be said about the intellectual background from which Watson arose and against which his contribution must be understood and appraised.

Among the relatively few men who had an appreciable impact on the development of ideas, even fewer, probably only a minority, were also able to see their own role in this development correctly. Watson does not belong to that doubly distinguished minority. Moreover, it would seem that those who up to now have written about him and about the recent history of psychology in this respect share his illusions. As Watson sees himself and, following him, as these writers see him, he is above all a radical innovator. Functionalism, accordingly, is merely the last gasp of classical psychology. Watson made the break; with him modern psychology began. As I see him, Watson is above all

a completer and a consummator—the greatest, though not chronologically the last, of the Functionalists. The break, accordingly, occurred between Structuralism and Functionalism. With the latter, modern psychology had begun before Watson appeared on the scene. The one methodological contribution of Watson which is specifically his own is merely a footnote—though, I insist, a most important one—to the methodological ideas of the Functionalists. These are somewhat unconventional assertions. Starting with them, I have once more, for clarity's sake, reversed the order of thesis and argument. So I hurry to supply the latter.

Experimentally, work of the kind the Functionalists emphasized had of course been done, at least incidentally, before their time. Systematically, though, they were, under the impact of Darwinism, the first to introduce certain ideas into the very groundwork—to introduce them, as it were, on the very ground floor of their conception of psychology. The most important of these ideas is that of process. According to the Functionalists, the psychologist must, like any other scientist, search for process laws. Specifically, this means that he must search for laws which permit him to predict, from sufficient information about the present state of a man, or any number of men, and their environment, the future course of events concerning both the man or men and the environment. To us this is a truism, except that, as we have since realized, we may have to supplement the information about the present with some about the past. Or, at least, we must be prepared to do that at certain stages of our knowledge. In the broad pattern which I am trying to trace, this qualification, though I think not unimportant, is merely a detail. Details apart, the emphasis on in principle comprehensive process is by now a truism. Let me say, then, as a simple matter of historical justice, that in psychology those who made it a truism are the Functionalists and Watson. The unconventional part of my thesis lies of course in the last two words: "and Watson." You will not on this occasion expect a full documentation for either its conventional or its unconventional part. But I shall adduce some facts which should at least lend some credence to both.

If I were allowed a choice of only three documents from which to teach the history of Functionalism I would choose, first, Dewey's "The Reflex Arc Concept in Psychology," of 1896, which is the opening shot of the movement; second, J. R. Angell's presidential address, "The Province of Functional Psychology," of 1907, which marks its first peak; and, third, Carr's text, *Psychology*, of 1925. There is, of course, a difference between these documents. The first two are what historians call primary sources; the third, being a text, is merely a secondary source; however, it was for its time a very good text; it was, I take it, very successful; and it represents very adequately the maturity of the movement. Now in Carr's text our great truism is not only stated precisely, without any flaw whatsoever, but it permeates the whole book. In Dewey's paper you will look for it in vain. (This is easily understood from Dewey's intellectual background, to which I could not possibly attend on this occasion.) In Angell's address you will find it, but only if you know how to look for it. If you do, then you will discover it here and there, but always, or virtually always, with a certain metaphysical blur. The blur stems, both logically and historically, from the metaphysical purposivism of the philosophical Darwinians. If you reflect on this temporal sequence extending through about three decades, from Dewey over Angell to Carr, you will, I hope, at

least find plausible what I believe to be a historical fact, namely, that within this development the first incisive statement of our great truism, made by a psychologist for psychologists, lucidly and without any teleological blur, occurs in Watson's classical "Psychology as the Behaviorist Views It," of 1913. This is what I had in mind when I called him the great consummator of Functionalism. With Carr, if I may so put it without disrespect, the insight had reached textbook level.

We are ready to see how all this squares with epiphenomenalism or, rather, with Watson's failure to adopt it and, as I said, let it go at that. In the paper just mentioned he accused the Functionalists of being interactionists who merely pose as epiphenomenalists when they want to give themselves the air of scientific righteousness. In the calm light of a later day the accusation seems grossly unfair. The Functionalists were not really interactionists. They were much too sane for that (always excepting Dewey, who is a case by himself). On the other hand, though, while the substance and structure of their thought was epiphenomenalist, they did not at first come right out and say so. Angell, in his politically very astute presidential address, even derided what he calls the pale ghost of epiphenomenalism. This helps to explain why Watson made those accusations. To understand fully one must, I think, recall the larger scheme of things, academic as well as nonacademic, intellectual as well as ideological, which the Functionalists found when, about half a generation before Watson, they opened their campaign. The then regnant philosophy was either dualistic or idealistic and, therefore, not at all favorable to a science of man.

Cleverly taking advantage of some favorable circumstances that were also present, e.g., the American enthusiasm for "applied science," the Functionalists tried to make in this scheme of things a place for psychology as we know it. Their success is a matter of record. Naturally, it required statesmanship and, perhaps, a modicum of diplomatic equivocation. (There are always some who sincerely believe what is also expedient.) This is how I would account for those occasional purple patches in their earlier writings, such as Angell's derision of epiphenomenalism or his quaintly Victorian hymn, in the same piece, to the "stellar role" mind plays in the "cosmic sweep" of evolution. Young Watson, alas, was neither statesman nor diplomat. I can easily imagine that he grew quite angry at all that rhetoric. Among the cruder ideological dynamisms, such indignation was, perhaps, the major cause for his rejection of even epiphenomenalism as not sufficiently radical. There were contributory causes, to be sure, such as the ideas of Loeb and Watson's own commitment to animal work. I shall pass these over and, finally, turn to his specific contribution, the one which for the sake of emphasis I called a footnote, though, I repeat, a very important one.

The contribution was not, as probably Watson thought, his materialism or *metaphysical behaviorism*—i.e., the thesis, which is merely silly, that there are no minds—but, rather, his *methodological behaviorism*. Let me then for a moment disregard history and present what I take to be a correct and defensible statement of this thesis. Consider three kinds of variables; call them behavioral, physiological, and environmental, respectively. Behavior and, therefore, the behavioral variables are physical. Smiling, frowning, talking, and so on, whatever else they may be or betoken, are certainly also physical events. The nature of the physiological variables I shall take for granted. They describe, for the most part though

not exclusively, the stuffings of the organism. The environment in which we move is to a very large extent social. This, however, means merely that it contains, in addition to other kinds of physical objects, other people. Thus the environment could in principle be described in terms of physical, physiological, and behavioral variables. Notice, incidentally, that I just used 'physical' in a narrower sense which is quite customary. Notice, furthermore, that I did not claim that these three kinds of variables are independent of each other. Combine, next, Watson's good thesis, that there are no interacting minds, with the great Functionalist truism about process. What follows is the thesis of methodological behaviorism. It must *in principle* be possible to predict future behavior, including verbal behavior, from a sufficiency of information about present (and past) behavioral, physiological, and environmental variables. This is the thesis. Let me also state one of its corollaries. Speaking commonsensically or, for that matter, clinically, we often attribute to a person a certain state of mind. We say that he or she perceives something, remembers something, plans something, is sad or gay, and so on. Nor is there any doubt, commonsensically, that we know what we mean when we assert such things, or, for that matter, that sometimes we are even right. It follows that it must *in principle* be possible to coordinate to any such statement another one, however complex, which mentions only behavioral, physiological, and environmental items, such that they are either both true or both false. Otherwise one would have to maintain that we can, literally and not metaphorically speaking, directly observe other people's states of mind. Few nowadays are either bold or foolish enough to assert that.

As I have just cautiously formulated it, methodological behaviorism, like Functionalism, has conquered itself to death. It, too, has become a truism. Virtually every American psychologist, whether he knows it or not, is nowadays a methodological behaviorist. That goes for those who glorify John B. Watson as well as for those who belittle him. He himself has at several places stated the thesis very clearly. But then he also said so much else that could be, and has been, misunderstood. Some comments are therefore in order.

Notice, first, in my formulation, the important qualification "in principle." The phrase is ambiguous. One thing we may mean when accepting something with this qualification is that it is not anything to be acted upon but, as one says, merely a frame of reference, logical or philosophical. The other thing we may mean is that we accept something as a program to be acted on and which, therefore, after a reasonable time may be judged by its fruits. Methodological behaviorism is a matter of principle in both senses; and it has, as I see it, in both senses so far stood the test of time. Philosophically, it is certainly not the whole of the best contemporary analysis of the traditional tangle known as the mind-body problem. But it fits into this analysis and is, in fact, an important part of it. This analysis as a whole avoids the many perplexities of traditional epiphenomenalism as well as those of Watsonian materialism. These, however, are matters better left to the epistemological seminar. So I shall say no more about them on this occasion. As a program, methodological behaviorism, as I just cautiously formulated it, has done well, too. To be overly modest is, to my mind, as unwise as it is to claim too much. Scientifically, psychologists do not know very much about human behavior, certainly not as much as they would like; but they by now do know

quite a bit. And, whatever they thus know they state, wittingly or unwittingly, in behavioristic terms. At least, they try to do that; they usually succeed to a considerable extent; and they are the better pleased the more nearly they succeed. Nor is it, I think, unreasonable to claim that this tendency has had a salutary and sobering effect and has thus helped us, at least indirectly, to acquire whatever little knowledge we have. This is the bright side of the picture. Let me now turn to another.

During the late thirties and early forties Watsonianism was ardently embraced by a group of then very vociferous philosophers, the so-called physicalists. The offspring of the embrace was *physicalistic behaviorism*, which is not at all the same thing as methodological behaviorism. A physicalist insists not only, as a methodological behaviorist does, that in the science of psychology we can and in principle must speak about such things as, say, smiling and frowning as physical events. He also insists that we can and in principle must avoid these very words, smiling and frowning. That is to say, he insists that in the science of psychology we can and in principle must describe a smiling or a frowning face without using any other words than those used in describing, literally and not metaphorically, inanimate objects such as a machine or a landscape. The difference is clearest where it is also most important, namely, in the case of language. Assume that we hear somebody say that he is sad or, for that matter, that he is cold. The methodological behaviorist assumes and, particularly in the area of personality, often must assume, not necessarily that the speaker is sad or cold, for he may of course be lying or attitudinizing, but, rather, that the noises "mean something" and, also, that the speaker "knows what they mean." Not so the physicalist. He merely hears noises. Thus he cannot, as a scientist, use the two phrases within quotation marks without first providing for them equivalents which contain only words belonging to his very limited vocabulary. This is what physicalistic behaviorism means. What shall we think of it?

Philosophically I think it is both sound and important. The reason for its being sound is that philosophical analysis is, to a very large extent and in a subtle way which I cannot take time to explain, a matter of principle only, in the sense in which this phrase does not signify a program to be acted upon. As a scientific program physicalistic behaviorism always was, still is, and for all I know may forever remain fantastically unrealistic. Methodological behaviorism is sound both as a frame of reference and as a program. The difference is worth pointing out. For, had it always been clearly understood, we might have saved ourselves much commotion: needless apprehensions and often quite irrational aversions by the clinical and social psychologists on the one hand, much shallow talk and a good deal of silly doctrinairism by the physicalists on the other. Under the circumstances it is natural to ask what stand Watson himself took on this issue or, perhaps better, pseudo-issue. The difficulty one encounters in trying to answer the question is that, when the issue came to the fore, Watson had already withdrawn from public participation in such debates. Thus the best one can do is to look for hints in his earlier writings. In the animal field, from which he came, the difference makes, of course, no difference. In his writings on personality he started, with eminent common sense and without any compunction whatsoever, from the variables of the clinical psychologist. Logically, in a

footnote to his very first methodological paper, the one mentioned before, he refers to what he there calls the "language method" as merely an expedient short cut. In some cases that is, indeed, all it is. In some others, the gap which the so-called short cut bridges is, logically speaking, very wide. One may well wonder whether Watson ever appreciated how wide it really is.

This brings me to my next point, no longer about the contribution as such and its later vicissitudes but, rather, about the historical Watson. The point once needed emphasis very badly in order to counteract one of the most widely spread and most shallow clichés. Even now such emphasis is not wholly lost. In order to provide it, I shall avail myself of a phrase that was dear to the very propagators of those clichés. Watson's psychology was not, at any stage of his thought, one of "muscle twitches." To say the same thing more academically, Watson never was and, I venture to guess, would not be today a doctrinaire advocate of physiological reduction. He was, quite to the contrary, a vigorous and, if anything, overoptimistic advocate of behavioral research, clinically as well as in the laboratory.

I have just touched on a contemporary issue. Such issues are tender to the touch. In order to reduce the sensitivity, I shall avail myself of one of the precautions I built into the formulation of methodological behaviorism. Remember that I listed three kinds of variables: behavioral, environmental, and physiological. If there are no interacting minds, as of course there are not, and if the world is completely and comprehensively lawful, as of course it is (in spite of the *apparent* exception of quantum mechanics, which is so frequently misunderstood but to which I cannot possibly attend on this occasion), then it follows that there are complete and comprehensive process laws covering human behavior, which contain, though of course not independently, at most our three kinds of variables. It does not follow that there are also such laws containing only two of the three kinds, namely, the behavioral and environmental ones. In other words, it may well be and it is, in fact, quite plausible that if we pursue purely behavioral research long enough, we shall *in the long run* come upon a barrier. That is to say, we may eventually reach a point where, though the laws we shall then have will not be complete and comprehensive process laws, we shall yet not be able to advance further as long as we limit ourselves to these two kinds of variables while, on the other hand, we could do better if we introduced the third kind, i.e., the physiological variables. This long-run possibility or, if you prefer, plausibility, is the sound core of what is in all other respects the purely ideological clamor of the contemporary enthusiasts for physiological reduction. I call the clamor ideological because, as I see it, its real purpose, whether conscious or unconscious, is not to make the very theoretical and quite uncontroversial point on which I have just insisted but, rather, to talk up the *short-run* prospects of research in physiological psychology by talking down the short-run prospects of behavioral work. Such talk judges itself. Behavioral research has produced at least some knowledge, however incomplete and imperfect in many respects it may be, in areas to which physiological research has as yet contributed nothing. In his own time Watson vigorously took the side of behavioral work. So, I ventured to guess, he would today. He even was, as I also mentioned, overoptimistic about both the short-run and the long-run prospects of purely behavioral work. The overoptimism fits well into the picture I have formed for myself of the man. Watson, as I see him,

was above all a man of action, primarily a doer and only incidentally, as it were, a thinker. Doer that he was, he was eager to start building the house of psychology. In this task, which he gravely underestimated, he knew that he could expect no help from the physiological psychologists of his day, just as behavioral research today can as yet expect but little help from these quarters. This, I believe, is one major reason why Watson was not in his day and probably would not be today what I call an ideological reductionist. (Nonideologically, there is no issue.) But there was still another such reason. Watson had an amazingly naive and almost superstitious distrust of any appeal to the action of the central nervous system. Reading him, one cannot help feeling that that evil thing, that darling of priests and kings, that enemy of all human happiness and progress, the interacting mind, begins one half of an inch beneath the skin of the skull. How queer! How amusing in one sense and how pathetic in another! But, then, we had better accept our great men as they are. Let me connect this comment with another.

Contemporary behaviorism, if there still is such a thing in addition to a group of important truisms, is certainly not another "school" of psychology. Early behaviorism or Watsonianism was the last of the schools; certainly it was more of a school than Functionalism in its maturity. (This neglects psychoanalysis, which is a historical phenomenon of different magnitude, as well as Gestalt, which, though chronologically later, belongs structurally to the German nineteenth century.) A school, as I use the term, is a group of men who propound a doctrine that is a mixture of three quite heterogeneous ingredients. (Also, a school is typically associated with the advocacy of, or, at least, emphasis on a special research interest.) The three heterogeneous ingredients of the classical mixture are, first, a more or less articulate philosophical position; second, some methodological insights *about* psychology; and, third, a group of beliefs concerning matters of fact and law *within* psychology. Such matters of fact and law are, of course, philosophically and methodologically quite neutral. The members of the school nonetheless fuse and confuse them with their philosophical and methodological tenets, as if they were a logically inseparable part of them. In Watson's case the first ingredient is his materialism or metaphysical behaviorism, the thesis that there are no minds. The second ingredient is his specific contribution, methodological behaviorism. The associated research interest was animal work. The third ingredient is his trio of pseudo-methodological tenets, as I like to call them. There is, first, his peripheralism. There is, second, his inclination to favor pure contiguity over so-called reinforcement theories of learning. There is, third, his denial of individual differences. The first two are related to his metaphysics. The third is very intimately connected with his social philosophy. I shall take them up in this order and, in conclusion, say a few words about his social philosophy.

Watson propounded, dogmatically and on ideological grounds, a peripheral theory of thought and a glandular theory of emotion. My emphasis is on the adverbial clause, "dogmatically and on ideological grounds"; for the facts, whatever they may be, are philosophically and methodologically neutral. This commitment was probably among the circumstances that saved Watson from ideological reductionism. The intellectual motive behind it appears clearest in his second fundamental paper, "Image and Affection in Behavior," of 1913, which he read before the Columbia seminar. A phrase of Külpe's provides

the cue. Külpe, as you may remember, distinguished between peripherally and centrally aroused sensations. What he meant is perfectly clear. An image I now have of my mother's face is not immediately preceded by the sort of peripheral stimulation I used to have when I still could look at her. Such an image is, according to Külpe, centrally aroused. That does not mean that it is outside the causal nexus of our three kinds of variables. Watson, however, confused 'centrally aroused' with 'uncaused.' Images, he concluded, are children of the devil, or, more precisely, of an interacting mind. Thus he decreed that there are none, except perhaps a sporadic few which do not matter. Strangely, he did not see that even a single one would suffice to destroy his case. If any one wonders whether the mistake is really that obvious, I suggest that he examine the text. Watson's peripheralism is beyond all doubt tied up with his antimentalism. Motivations, though, are complex. His activism probably had its part in this preference. For, surely, he realized that in the eventual development of physiological psychology, detailed knowledge and control of the central nervous system would in all probability come last.

The glands, which according to this early paper are the peripheral seat of the emotions, are the sex glands. Nor does Watson miss the opportunity to endorse some of the basic ideas of psychoanalysis. Naturally, he objected and in his later writings continued to object vigorously to its mentalistic bias. Yet he recognized its significance as early as 1913. What appealed to him above all was the insight, which he himself elaborated in his later writings, that man's behavior, however complex, can and must be causally explained in terms of his primary needs. The first to build an elaborate structure on this ground was Freud. As far as I know, the first experimental psychologist who understood him and had the courage to speak up for him was Watson. In this respect his record is admirable indeed. Freud's remarks about behaviorism in his *Autobiography* of 1925 leave no doubt that he, for all his towering greatness, did not understand Watson.

The emphasis on learning, important and influential as it was, merely continued what the Functionalists had auspiciously begun. The distrust of reinforcement theories which I attributed to Watson was his own. Notice, though, that I was cautious, speaking merely of an inclination to favor what we now call contiguity theories. The reason for my caution is that, when Watson was active, theorizing in this area was as yet most rudimentary. So it does not seem fair to interpret anything he then said as a firm commitment to any of the more elaborate theoretical structures of today. The ideological root of whatever preference for contiguity he had is not hard to discover. In an understandable—though, again, not very subtle—reaction to speculative Darwinism, he probably saw reinforcement theories as an entering wedge for metaphysical purposivism. By now it is understood that there is no logical connection between the two. Foremost among the psychologists who in the twenties and early thirties clarified this point is probably Tolman. The philosopher from whom he received support is R. B. Perry.

The third pseudo-methodological tenet is an extreme environmentalism. I shall first consider it logically, then as the expression of a social gospel. According to Watson, a child is at birth equipped with nothing but the structure of its body and a few elementary unlearned responses. All the rest is learning. What that means depends on what is meant by 'structure.' If structure includes fine structure, including all the as yet unknown characteristics of the

central nervous system, then the thesis is certainly true but, alas, only because it says virtually nothing. Certainly it is broad enough to allow for the inheritance of very specific talents, and even personality traits. If, on the other hand, 'structure' merely means gross structure, or relatively gross structure, which is what Watson meant, then his pronouncement was and, for all I know, still is most dogmatic. To me, it is not even plausible. So I shall look at the gospel behind the dogma.

The best way and, if one must be brief, the only way to describe a man's social philosophy is to describe his Utopia, i.e., the society he considers ideal. Watson's Utopia has been very accurately described in two novels, Aldous Huxley's *Brave New World* and Skinner's *Walden Two*. They leave no doubt that he stands in the succession of Plato. The one major difference is that while Plato wants us to be ruled by philosopher-kings, Watson's chosen ruler is the scientist and, in particular, the psychologist. His Utopia is thus a totalitarian technocracy. I use the word technocracy deliberately. It indicates, at least succinctly, how Watson fits into the temper of his time. The other great intellectual technocrat of his day, though by far not as totalitarian, was Thorstein Veblen. His chosen rulers were the engineers. The bridge of wishful thinking which, in Watson's case, leads from his Utopia to his environmentalism is rather obvious. Man is infinitely plastic and thus *can* be molded because he *ought* to be molded.

This is not the place to review those features of the American scene during the first quarter of the century which make it plausible that such men as Watson and Veblen held these views. Nor do I wish to make any dogmatic pronouncements of my own on such difficult questions as the role of the expert in a technological society, or on how democratic ideals can be reconciled with a sober appraisal of human nature. In his admirable piece on the Moses of Michelangelo, Freud described the emotion—a mixture of compassion, anger, and indignation—which sometimes overcomes an outstanding man when he reflects on the misery of the human condition. Watson, I like to think, was overwhelmed by this mixed emotion when he wrote, as unhappily he did, particularly in the last chapter of his *Behaviorism*, of 1925, as if he were the sworn enemy of freedom and the inner life. So I shall limit myself to a few words about the intellectual confusions which were among the causes of these unfortunate statements.

Social arrangements differ significantly in the number and kind of choices they provide. Some choices, such as that of a profession or a mate, are "important"; some others are less so. The consequences, including all sorts and degrees of "penalties," which by social arrangement attend a choice, are of course part of the "choice situation." Along such lines a notion of human freedom can be elaborated. Perhaps it is vague. It may even be irremediably vague. That does not mean that it is either "metaphysical" or incompatible with the idea of comprehensive process. Watson did not grasp the distinction. For him, freedom, in any sense, was an illusion spawned by that arch-evil, the illusion of interacting minds. So he wrote almost as if he had set out to provide ammunition for those who now try to convince us that if we want to preserve human freedom, we must abandon the idea of a science of man, and, instead, adopt their creeds. What goes for freedom goes equally for what I called the inner life. For what I mean by this phrase, the inner life, is nothing "metaphysical," but rather the availability, the complexity, the originality, and the adequacy of the implicit verbal response. Or, at

least, that is what it amounts to in psychological terms. Watson, once more, failed to grasp the distinction.

I do not feel like ending on this sour note. Some painters overemphasize both the lights and the shadows, particularly when they are asked to do a portrait in a very limited time. Perhaps I have done the same in trying to sketch an intellectual portrait of John B. Watson. Yet I have not the slightest doubt that, with all the light and all the shadow, he is a very major figure. Psychology owes him much. His place in the history of our civilization is not inconsiderable and it is secure. Such men are exceedingly rare. We ought to accept them and appreciate them for what they are.

(Received November 18, 1955)

• Perception and the Representative Design of Psychological Experiments

By Egon Brunswik

Definitive presentation of Brunswik's pioneer work in experimental psychology. Shows how criteria of representativeness can be incorporated into experimental design, applying probabilistic sampling to both subjects and objects. Analysis of the experimental literature focuses on a general theory of adaptive behavior, with special emphasis on the cognitive functions of the organism.

166 pages, 39 figs. $5.00

• Nonverbal Communication

NOTES ON THE VISUAL PERCEPTION OF HUMAN RELATIONS

By Jurgen Ruesch, M.D., and Weldon Kees

Analyzes little-explored aspects of interpersonal communication by means of numerous photographs, succinct captions, and text. Of fundamental importance to psychologists, and to workers in psychiatry, art, advertising, decoration, photography, and sociology.

216 pages, 330 illus. $7.50

UNIVERSITY OF CALIFORNIA PRESS

Address: **Berkeley 4, California**